## REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-98-

0800

38

Public reporting burden for this collection of information is estimated to average 1 hour per response, in ırces, gathering
and maintaining the data needed, and completing and reviewing the collection of information. Send ıis collection of
information, including suggestions for reducing this burden, to Washington Headquarters Services, Direc Highway, Suite
1204, Arlington, VA 22202-4302, and to the Office of management and Budget, Paperwork Reduction Prc

| | | |
|---|---|---|
| 1. AGENCY USE ONLY (Leave Blank) | 2. REPORT DATE<br>November, 1994 | 3. F<br>Final |

4. TITLE AND SUBTITLE
USAF Summer Research Program - 1993 Summer Research Extension Program Final Reports, Volume 4B, Wright Laboratory

5. FUNDING NUMBERS

6. AUTHORS
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research and Development Labs, Culver City, CA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
4040 Fairfax Dr, Suite 500
Arlington, VA 22203-1613

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Contract Number: F4962-90-C-0076

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this program is to develop the basis for continuing research of interest to the Air Force at the institution of the faculty member; to stimulate continuing relations among faculty members and professional peers in the Air Force to enhance the research interests and capabilities of scientific and engineering educators; and to provide follow-on funding for research of particular promise that was started at an Air Force laboratory under the Summer Faculty Research Program. Each participant provided a report of their research, and these reports are consolidated into this annual report.

14. SUBJECT TERMS
AIR FORCE RESEARCH, AIR FORCE, ENGINEERING, LABORATORIES, REPORTS, UNIVERSITIES

15. NUMBER OF PAGES

16. PRICE CODE

| | | | |
|---|---|---|---|
| 17. SECURITY CLASSIFICATION OF REPORT<br>Unclassified | 18. SECURITY CLASSIFICATION OF THIS PAGE<br>Unclassified | 19. SECURITY CLASSIFICATION OF ABSTRACT<br>Unclassified | 20. LIMITATION OF ABSTRACT<br>UL |

# UNITED STATES AIR FORCE

# SUMMER RESEARCH PROGRAM -- 1993

# SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

## VOLUME 4B

## WRIGHT LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES
5800 Uplander Way
Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major David Hart

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Gwendolyn Smith

Program Administrator, RDL
Johnetta Thompson

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling Air Force Base
Washington, D.C.
November 1994

# PREFACE

This volume is part of a five-volume set that summarizes the research of participants in the 1993 AFOSR Summer Research Extension Program (SREP). The current volume, Volume 4B of 5, presents the final reports of SREP participants at Wright Laborarory.

Reports presented in this volume are arranged alphabetically by author and are numbered consecutively -- e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3, with each series of reports preceded by a 35 page management summary. Reports in the five-volume set are organized as follows:

| VOLUME | TITLE |
|---|---|
| 1A | Armstrong Laboratory (part one) |
| 1B | Armstrong Laboratory (part two) |
| 2 | Phillips Laboratory |
| 3 | Rome Laboratory |
| 4A | Wright Laboratory (part one) |
| 4B | Wright Laboratory (part two) |
| 5 | Arnold Engineering Development Center<br>Frank J. Seiler Research Laboratory<br>Wilford Hall Medical Center |

# 1993 SREP FINAL REPORTS

## Armstrong Laboratory

## VOLUME 1A

# 1993 SREP FINAL REPORTS

## Armstrong Laboratory

## VOLUME 1B

# 1993 SREP FINAL REPORTS

## Phillips Laboratory

## VOLUME 2

## 1993 SREP FINAL REPORTS

## Phillips Laboratory

## VOLUME 2
cont'd

# 1993 SREP FINAL REPORTS

## Rome Laboratory

## VOLUME 3

# 1993 SREP FINAL REPORTS

## Wright Laboratory

## VOLUME 4A

# 1993 SREP FINAL REPORTS

## Wright Laboratory

## VOLUME 4A
## cont'd

# 1993 SREP FINAL REPORTS

## Wright Laboratory

## VOLUME 4B

# 1993 SREP FINAL REPORTS

## Wright Laboratory

## VOLUME 4B
## cont'd

# 1993 SREP FINAL REPORTS

## VOLUME 5

### Arnold Engineering Development Center

### Frank J Seiler Research Laboratory

### Wilford Hall Medical Center

# 1993 SUMMER RESEARCH EXTENSION PROGRAM (SREP) MANAGEMENT REPORT

## 1.0 BACKGROUND

Under the provisions of Air Force Office of Scientific Research (AFOSR) contract F49620-90-C-0076, September 1990, Research & Development Laboratories (RDL), an 8(a) contractor in Culver City, CA, manages AFOSR's Summer Research Program. This report is issued in partial fulfillment of that contract (CLIN 0003AC).

The Summer Research Extension Program (SREP) is one of four programs AFOSR manages under the Summer Research Program. The Summer Faculty Research Program (SFRP) and the Graduate Student Research Program (GSRP) place college-level research associates in Air Force research laboratories around the United States for 8 to 12 weeks of research with Air Force scientists. The High School Apprenticeship Program (HSAP) is the fourth element of the Summer Research Program, allowing promising mathematics and science students to spend two months of their summer vacations working at Air Force laboratories within commuting distance from their homes.

SFRP associates and exceptional GSRP associates are encouraged, at the end of their summer tours, to write proposals to extend their summer research during the following calendar year at their home institutions. AFOSR provides funds adequate to pay for 75 SREP subcontracts. In addition, AFOSR has traditionally provided further funding, when available, to pay for additional SREP proposals, including those submitted by associates from Historically Black Colleges and Universities (HBCUs) and Minority Institutions (MIs). Finally, laboratories may transfer internal funds to AFOSR to fund additional SREPs. Ultimately the laboratories inform RDL of their SREP choices, RDL gets AFOSR approval, and RDL forwards a subcontract to the institution where the SREP associate is employed. The subcontract (see Appendix 1 for a sample) cites the SREP associate as the principal investigator and requires submission of a report at the end of the subcontract period.

Institutions are encouraged to share costs of the SREP research, and many do so. The most common cost-sharing arrangement is reduction in the overhead, fringes, or administrative charges institutions would normally add on to the principal investigator's or research associate's labor. Some institutions also provide other support (e.g., computer run time, administrative assistance, facilities and equipment or research assistants) at reduced or no cost.

When RDL receives the signed subcontract, we fund the effort initially by providing 90% of the subcontract amount to the institution (normally $18,000 for a $20,000 SREP). When we receive the end-of-research report, we evaluate it administratively and send a copy to the laboratory for a technical evaluation. When the laboratory notifies us the SREP report is acceptable, we release the remaining funds to the institution.

## 2.0 THE 1993 SREP PROGRAM

SELECTION DATA: A total of 719 faculty members (SFRP Associates) and 286 graduate students (GSRP associates) applied to participate in the 1992 Summer Research Program. From these applicants 185 SFRPs and 121 GSRPs were selected. The education level of those selected was as follows:

| 1992 SRP Associates, by Degree | | | |
|---|---|---|---|
| SFRP | | GSRP | |
| PHD | MS | MS | BS |
| 179 | 6 | 52 | 69 |

Of the participants in the 1992 Summer Research Program 90 percent of SFRPs and 25 percent of GSRPs submitted proposals for the SREP. Ninety proposals from SFRPs and ten from GSRPs were selected for funding, which equates to a selection rate of 54% of the SFRP proposals and of 34% for GSRP proposals.

| 1993 SREP: Proposals Submitted vs. Proposals Selected | | | |
|---|---|---|---|
| | Summer 1992 Participants | Submitted SREP Proposals | SREPs Funded |
| SFRP | 185 | 167 | 90 |
| GSRP | 121 | 29 | 10 |
| TOTAL | 306 | 196 | 100 |

The funding was provided as follows:

| | |
|---|---|
| Contractual slots funded by AFOSR | 75 |
| Laboratory funded | 14 |
| Additional funding from AFOSR | 11 |
| Total | 100 |

Six HBCU/MI associates from the 1992 summer program submitted SREP proposals; six were selected (none were lab-funded; all were funded by additional AFOSR funds).

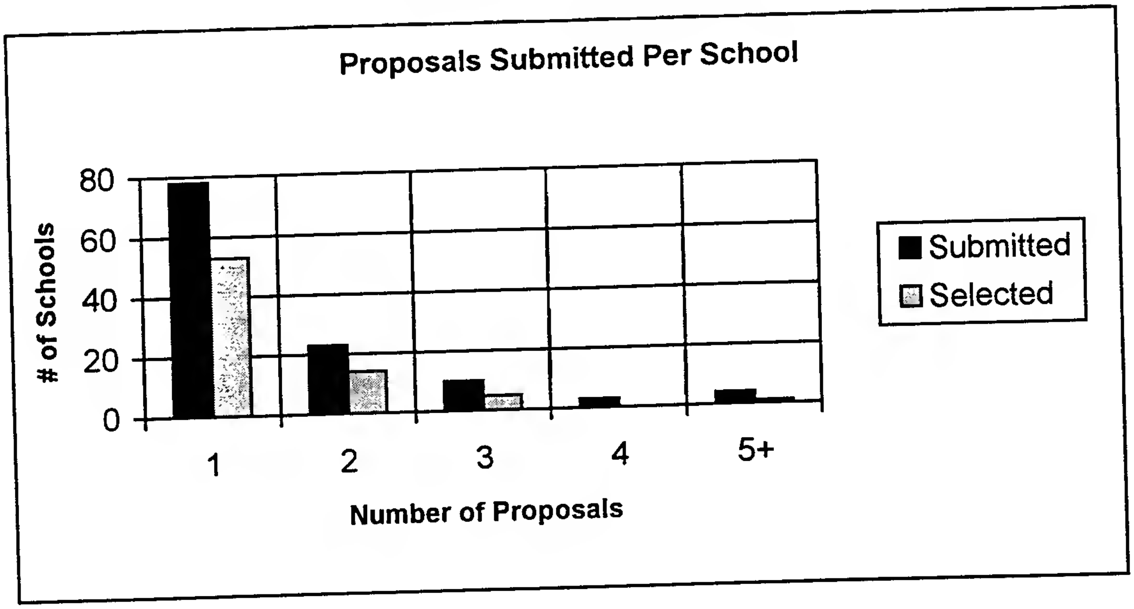
| Proposals Submitted and Selected, by Laboratory | | |
|---|---|---|
| | Applied | Selected |
| Air Force Civil Engineering Laboratory | 9 | 4 |
| Armstrong Laboratory | 41 | 19 |
| Arnold Engineering Development Center | 12 | 4 |
| Frank J. Seiler Research Laboratory | 6 | 3 |
| Phillips Laboratory | 33 | 19 |
| Rome Laboratory | 31 | 13 |
| Wilford Hall Medical Center | 2 | 1 |
| Wright Laboratory | 62 | 37 |
| TOTAL | 196 | 100 |

Note: Phillips Laboratory funded 3 SREPs; Wright Laboratory funded 11; and AFOSR funded 11 beyond its contractual 75.

The 306 1992 Summer Research Program participants represented 135 institutions.

| Institutions Represented on the 1992 SRP and 1993 SREP | | |
|---|---|---|
| Number of schools represented in the Summer 92 Program | Number of schools represented in submitted proposals | Number of schools represented in Funded Proposals |
| 135 | 118 | 73 |

Forty schools had more than one participant submitting proposals.



The selection rate for the 78 schools submitting 1 proposal (68%) was better than those submitting 2 proposals (61%), 3 proposals (50%), 4 proposals (0%) or 5+ proposals (25%). The 4 schools that submitted 5+ proposals accounted for 30 (15%) of the 196 proposals submitted.

Of the 196 proposals submitted, 159 offered institution cost sharing. Of the funded proposals which offered cost sharing, the minimum cost share was $1000.00, the maximum was $68,000.00 with an average cost share of $12,016.00.

| Proposals and Institution Cost Sharing | | |
|---|---|---|
| | Proposals Submitted | Proposals Funded |
| With cost sharing | 159 | 82 |
| Without cost sharing | 37 | 18 |
| Total | 196 | 100 |

The SREP participants were residents of 41 different states. Number of states represented at each laboratory were:

| States Represented, by Proposals Submitted/Selected per Laboratory | | |
|---|---|---|
| | Proposals Submitted | Proposals Funded |
| Air Force Civil Engineering Laboratory | 8 | 4 |
| Armstrong Laboratory | 21 | 13 |
| Arnold Engineering Development Center | 5 | 2 |
| Frank J. Seiler Research Laboratory | 5 | 3 |
| Phillips Laboratory | 16 | 14 |
| Rome Laboratory | 14 | 7 |
| Wilford Hall Medical Center | 2 | 1 |
| Wright Laboratory | 24 | 20 |

Eleven of the 1993 SREP Principal Investigators also participated in the 1992 SREP.

ADMINISTRATIVE EVALUATION: The administrative quality of the SREP associates' final reports was satisfactory. Most complied with the formatting and other instructions provided to them by RDL. Ninety seven final reports and two interim reports have been received and are included in this report. The subcontracts were funded by $1,991,623.00 of Air Force money. Institution cost sharing totaled $985,353.00.

TECHNICAL EVALUATION: The form used for the technical evaluation is provided as Appendix 2. ninety-two evaluation reports were received. Participants by laboratory versus evaluations submitted is shown below:

| | Participants | Evaluations | Percent |
|---|---|---|---|
| Air Force Civil Engineering Laboratory | * | * | * |
| Armstrong Laboratory | 23[1] | 20 | 95.2 |
| Arnold Engineering Development Center | 4 | 4 | 100 |
| Frank J. Seiler Research Laboratory | 3 | 3 | 100 |
| Phillips Laboratory | 19[2] | 18 | 100 |
| Rome Laboratory | 13 | 13 | 100 |
| Wilford Hall Medical Center | 1 | 1 | 100 |
| Wright Laboratory | 37 | 34 | 91.9 |
| Total | 100[3] | 93 | 95.9 |

*AFCEL was combined with Wright Laboratory's Flight Dynamics Directorate and Armstrong Laboratories Environics Directorate in 1993. All four of AFCEL's SREP awards went to Armstrong Laboratories Environics Directorate, and their reports are included with Armstrong Lab.

Notes:

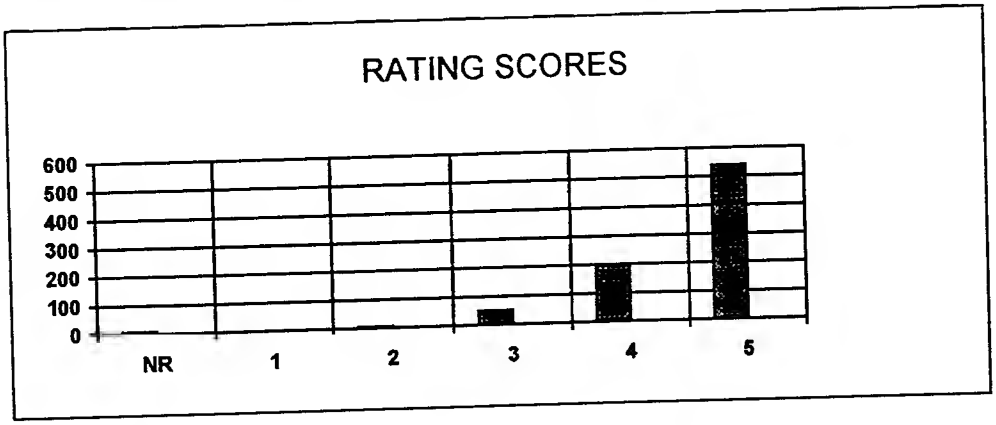
1: Research on two of the final reports was incomplete as of press time so there aren't any technical evaluations on them to process, yet. Percent complete is based upon 20/21=95.2%

2: One technical evaluation was not completed because one of the final reports was incomplete as of press time. Percent complete is based upon 18/18=100%

3: See notes 1 and 2 above. Percent complete is based upon 93/97=95.9%

The number of evaluations submitted for the 1993 SREP (95.9%) shows a marked improvement over the 1992 SREP submittals (65%).

PROGRAM EVALUATION: Each laboratory focal point evaluated ten areas (see Appendix 2) with a rating from one (lowest) to five (highest). The distribution of ratings was as follows:

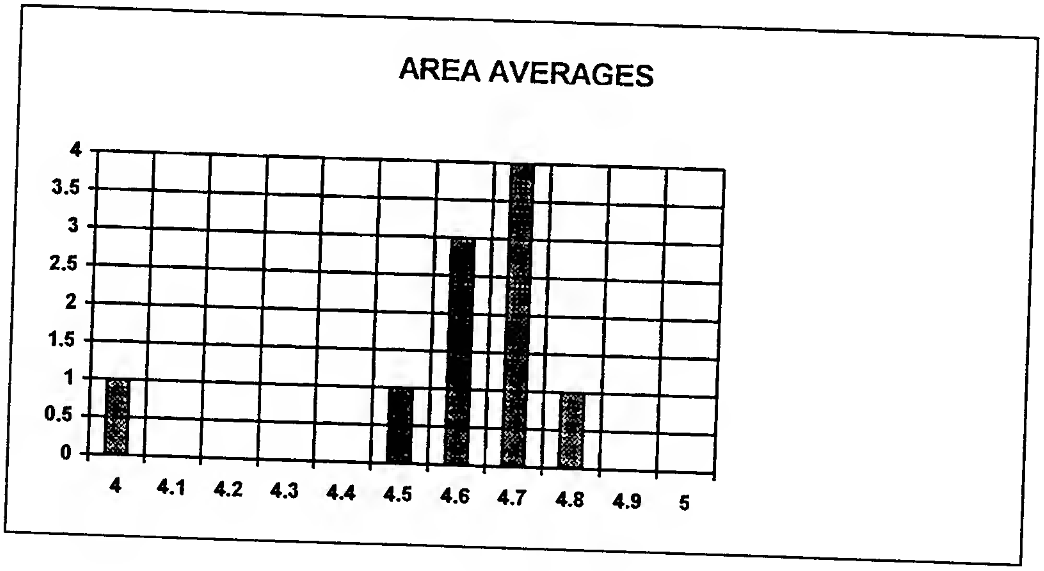


| Rating | Not Rated | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| # Responses | 7 | 1 | 7 | 62 (6%) | 226 (25%) | 617 (67%) |

The 8 low ratings (one 1 and seven 2's ) were for question 5 (one 2) "The USAF should continue to pursue the research in this SREP report" and question 10 (one 1 and six 2's) "The one-year period for complete SREP research is about right", in addition over 30% of the threes (20 of 62) were for question ten. The average rating by question was:

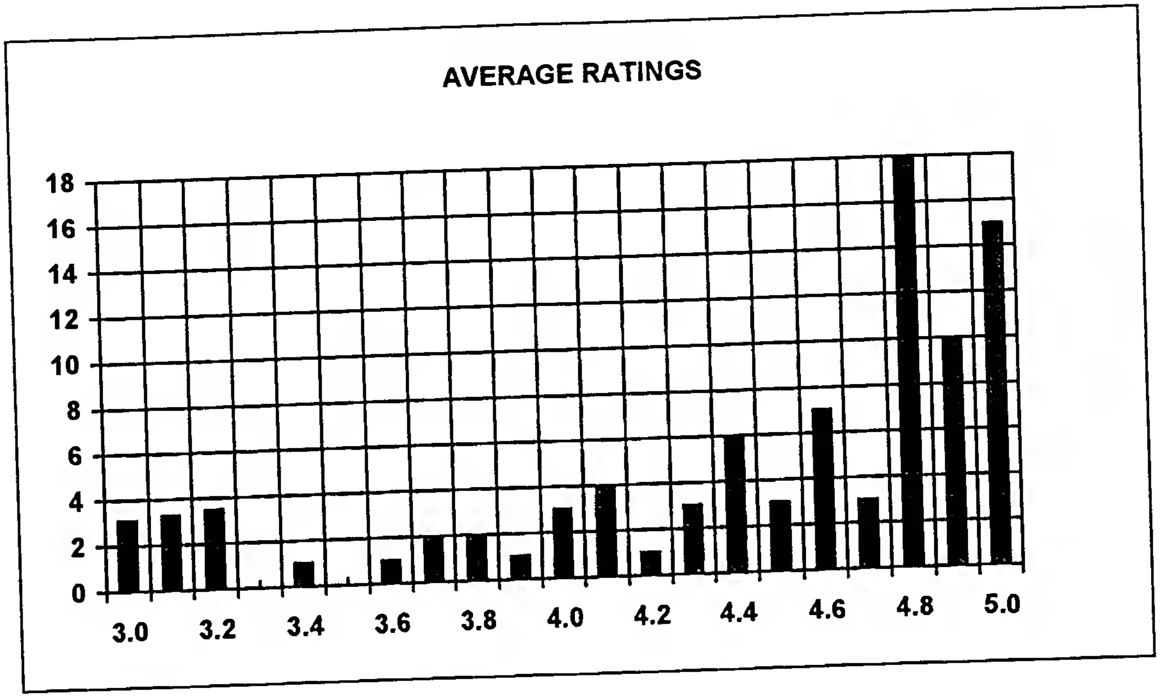
| Question | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Average | 4.6 | 4.6 | 4.7 | 4.7 | 4.6 | 4.7 | 4.8 | 4.5 | 4.6 | 4.0 |

The distribution of the averages was:



Area 10 "the one-year period for complete SREP research is about right" had the lowest average rating (4.1). The overall average across all factors was 4.6 with a small sample standard deviation of 0.2. The average rating for area 10 (4.1) is approximately three sigma lower than the overall average (4.6) indicating that a significant number of the evaluators feel that a period of other than one year should be available for complete SREP research.

The average ratings ranged from 3.4 to 5.0. The overall average for those reports that were evaluated was 4.6. Since the distribution of the ratings is not a normal distribution the average of 4.6 is misleading. In fact over half of the reports received an average rating of 4.8 or higher. The distribution of the average report ratings is as shown:



It is clear from the high ratings that the laboratories place a high value on AFOSR's Summer Research Extension Programs.

## 3.0 SUBCONTRACTS SUMMARY

Table 1 provides a summary of the SREP subcontracts. The individual reports are published in volumes as shown:

| Laboratory | Volume |
|---|---|
| Air Force Civil Engineering Laboratory | * |
| Armstrong Laboratory | 1 |
| Arnold Engineering Development Center | 5 |
| Frank J. Seiler Research Laboratory | 5 |
| Phillips Laboratory | 2 |
| Rome Laboratory | 3 |
| Wilford Hall Medical Center | 5 |
| Wright Laboratory | 4A, 4B |

*AFCEL was combined with Wright Laboratory's Flight Dynamics Directorate and Armstrong Laboratories Environics Directorate in 1993. All four of AFCEL's SREP awards went to Armstrong Laboratories Environics Directorate, and their reports are included with Armstrong Lab.

# 1993 SREP SUB-CONTRACT DATA

## TABLE 1: SUBCONTRACTS SUMMARY

| Report Author<br>Author's University | Author's Degree | Sponsoring<br>Lab | Performance Period | | Contract Amount<br>Univ. Cost Share |
|---|---|---|---|---|---|
| Abbott , Ben<br>Electrical Engineering<br>Vanderbilt University, Nashville, TN | M.S. | AEDC/ | 01/01/93 | 12/31/93 | $19619.00<br>$0.00 |
| Alberts , Thomas<br>Mechanical Engineering<br>Old Dominion University, Norfolk, VA | PhD | FJSRL/ | 01/01/93 | 04/15/94 | $20000.00<br>$8000.00 |
| Avula , Xavier<br>Mechanical & Aerospace Engineering<br>University of Missouri, Rolla, MO | PhD | AL/AO | 01/01/93 | 04/15/94 | $20000.00<br>$1836.00 |
| Balakrishan , S.<br>Mechanical & Aerospace Engineering<br>University of Missouri, Rolla, MO | PhD | WL/MN | 12/01/92 | 12/14/93 | $20000.00<br>$3996.00 |
| Baumgarten , Joseph<br>Mechanical Engineering<br>Iowa State University, Ames, IA | PhD | PL/VT | 01/01/93 | 04/01/94 | $19916.00<br>$9083.00 |
| Bayard , Jean-Pierre<br>Electrical & Electronic Engineering<br>California State University, Sacramento, CA | PhD | RL/ER | 01/01/93 | 12/31/93 | $20000.00<br>$7423.00 |
| Bellem , Raymond<br>Electrical & Computer Engineering<br>University of Arizona, Tucson, AZ | PhD | PL/VT | 01/01/93 | 02/28/94 | $19956.00<br>$0.00 |
| Biegl , Csaba<br>Electrical Engineering<br>Vanderbilt University, Nashville, TN | PhD | AEDC/ | 01/01/93 | 12/31/93 | $19999.00<br>$0.00 |
| Biggs , Albert<br>Electrical Engineering<br>University of Alabama, Huntsville, AL | PhD | PL/WS | 01/01/93 | 12/31/93 | $19975.00<br>$0.00 |
| Burkey , Theodore<br>Chemistry<br>Memphis State University, Memphis, TN | PhD | WL/MN | 01/01/93 | 12/31/93 | $20000.00<br>$18648.00 |
| Burkholder , Robert<br>Electrical Engineering<br>Ohio State University, Columbus, OH | PhD | WL/AA | 01/01/93 | 12/31/93 | $20000.00<br>$6727.00 |
| Callens, Jr. , Eugene<br>Mechanical and Industrial Engineer<br>Louisiana Tech University, Ruston, LA | PhD | WL/MN | 01/01/93 | 12/31/93 | $20000.00<br>$5700.00 |
| Chapman , Gary<br>Mechnical Engineering<br>University of California, Berkeley, CA | PhD | WL/MN | 01/01/93 | 12/31/94 | $20000.00<br>$0.00 |
| Chen , Chien-In<br>Electrical Engineering<br>Wright State University, Dayton, OH | PhD | WL/EL | 01/01/93 | 12/31/93 | $20000.00<br>$32065.00 |
| Chen , Jer-sen<br>Computer Science & Engineering<br>Wright State University, Dayton, OH | PhD | AL/CF | 01/01/93 | 12/31/93 | $20000.00<br>$31763.00 |

# 1993 SREP SUB-CONTRACT DATA

| Report Author<br>Author's University | Author's Degree | Sponsoring<br>Lab | Performance Period | Contract Amou<br>Univ. Cost Sha |
|---|---|---|---|---|
| Chen , Pinyuen<br>Mathematics<br>Syracuse University, Syracuse, NY | PhD | RL/IR | 01/01/93 12/31/93 | $20000.00<br>$0.00 |
| Chow , Joe<br>Industrial and Systems Engineering<br>Florida International University, Miami, FL | PhD | WL/FI | 01/01/93 01/14/94 | $20000.00<br>$2500.00 |
| Christensen , Thomas<br>Physics<br>University of Colorado, Colorado Springs, CO | PhD | FJSRL/ | 01/01/93 12/31/93 | $20000.00<br>$5390.00 |
| Collicott , Steven<br>Aeronautics and Astronautical Engineering<br>Purdue University, West Lafayette, IN | PhD | WL/FI | 01/01/93 12/31/93 | $20000.00<br>$13307.00 |
| Cooke , Nancy<br>Psychology<br>New Mexico State University, Las Cruces, NM | PhD | AL/HR | 01/01/93 12/31/93 | $20000.00<br>$6178.00 |
| Daley , Michael<br>Electrical Engineering<br>Memphis State, Memphis, TN | PhD | WHMC/ | 01/01/93 12/31/93 | $20000.00<br>$18260.00 |
| Davidson , Jennifer<br>Electrical Engineering<br>Iowa State University, Ames, IA | PhD | WL/MN | 01/01/93 02/28/94 | $19999.00<br>$0.00 |
| Deivanayagam , Subramaniam<br>Industrial Engineering<br>Tennessee Technological University, Cookeville, TN | PhD | AL/HR | 02/01/93 12/31/93 | $20000.00<br>$12491.00 |
| Elliott , David<br>Engineering<br>Arkansas Technology University, Russellville, AR | PhD | PL/RK | 10/01/92 08/15/93 | $20000.00<br>$50271.00 |
| Erdman , Paul<br>Physics and Astronomy<br>University of Iowa, Iowa City, IA | M.S. | PL/RK | 01/01/93 12/31/93 | $20000.00<br>$26408.00 |
| Ervin , Jamie<br>Mechanical and Aerospace Engineering<br>University of Dayton, Dayton, OH | PhD | WL/ML | 01/01/93 12/31/93 | $18632.00<br>$3000.00 |
| Erwin , Daniel<br>Aerospace Engineering<br>University of Southern California, Los Angeles, CA | PhD | PL/RK | 01/01/93 12/31/93 | $19962.00<br>$12696.00 |
| Ewert , Dan<br>Electrical Engineering<br>North Dakota State University, Fargo, ND | PhD | AL/AO | 01/01/93 12/31/93 | $20000.00<br>$2100.00 |
| Farmer , Barry<br>Materials Science and Engineering<br>University of Virginia, Charlottesville, VA | PhD | WL/ML | 01/01/93 02/28/94 | $20000.00<br>$2000.00 |
| Foo , Simon<br>Electrical Engineering<br>Florida State University, Tallahassee, FL | PhD | WL/MN | 01/01/93 12/31/93 | $19977.00<br>$0.00 |

# 1993 SREP SUB-CONTRACT DATA

| Report Author<br>Author's University | Author's Degree | Sponsoring Lab | Performance Period | | Contract Amount<br>Univ. Cost Share |
|---|---|---|---|---|---|
| Friedman , Jeffrey<br>Physics<br>University of Puerto Rico, Mayaguez, PR | PhD | PL/GP | 01/01/93 | 12/31/93 | $20000.00<br>$10233.00 |
| Gerstman , Bernard<br>Physics<br>Florida International University, Miami, FL | PhD | AL/OE | 01/01/93 | 04/30/94 | $19947.00<br>$2443.00 |
| Gillman , Ann<br>Chemistry<br>Eastern Kentucky University, Richmond, KY | M.S. | WL/PO | 01/01/93 | 12/31/93 | $20000.00<br>$15618.00 |
| Gimmestad , Gary<br>Research Institute<br>Georgia Institute of Technology, Atlanta, GA | PhD | PL/LI | 01/01/93 | 12/31/93 | $20000.00<br>$0.00 |
| Gould , Richard<br>Mechanical and Aerospace Engineering<br>North Carolina State University, Raleigh, NC | PhD | WL/PO | 01/01/93 | 12/31/93 | $20000.00<br>$8004.00 |
| Graham , Gary<br>Mechanical Engineering<br>Ohio University, Athens, OH | PhD | WL/FI | 01/01/93 | 12/31/93 | $20000.00<br>$5497.00 |
| Gramoll , Kurt<br>Aerospace Engineering<br>Georgia Institute of Technology, Atlanta, GA | PhD | AEDC/ | 01/01/93 | 12/31/93 | $19707.00<br>$14552.00 |
| Graul , Susan<br>Chemistry<br>Carnegie Mellon University, Pittsburgh, PA | PhD | PL/WS | 01/01/93 | 03/31/94 | $20000.00<br>$0.00 |
| Griffin , Steven<br>Engineering<br>University of Texas, San Antonio, TX | M.S. | PL/VT | 01/01/93 | 12/31/93 | $20000.00<br>$0.00 |
| Grubbs , Elmer<br>Electrical Engineering<br>New Mexico Highlands University, Las Vegas, NM | PhD | WL/AA | 01/01/93 | 12/31/93 | $20000.00<br>$6747.00 |
| Gupta , Pushpa<br>Mathematics<br>University of Maine, Orono, ME | PhD | AL/AO | 01/01/93 | 12/31/93 | $20000.00<br>$1472.00 |
| Hall , Ian<br>Materials Science<br>University of Delaware, Newark, DE | PhD | WL/ML | 01/01/93 | 12/31/93 | $20000.00<br>$9580.00 |
| Hedman , Paul<br>Chemical Engineering<br>Brigham Young University, Provo, UT | PhD | WL/PO | 01/01/93 | 12/31/93 | $19999.00<br>$7755.00 |
| Henry , Robert<br>Electrical & Computer Engineering<br>University of Southwestern Louisiana, Lafayette, LA | PhD | RL/C3 | 12/01/92 | 05/31/93 | $19883.00<br>$11404.00 |
| Henson , James<br>Electrical Engineering<br>University of Nevada, Reno, NV | PhD | PL/WS | 01/01/93 | 12/31/93 | $19913.00<br>$9338.00 |

# 1993 SREP SUB-CONTRACT DATA

| Report Author<br>Author's University | Author's Degree | Sponsoring<br>Lab | Performance Period | Contract Amou<br>Univ. Cost Sha |
|---|---|---|---|---|
| Hoe , Benjamin<br>Electrical Engineering<br>Polytechnic University, Brooklyn, NY | M.S. | RL/C3 | 09/01/92 05/31/93 | $19988.00<br>$7150.00 |
| Hughes , Rod<br>Psychology<br>Bowling Green State University, Bowling Green, OH | M.S. | AL/CF | 01/01/93 04/15/94 | $20000.00<br>$20846.00 |
| Hui , David<br>Mechanical Engineering<br>University of New Orleans, New Orleans, LA | PhD | WL/FI | 01/01/93 12/31/93 | $20000.00<br>$0.00 |
| Humi , Mayer<br>Mathematics<br>Worcester Polytechnic Institut, Worcester, MA | PhD | PL/LI | 01/01/93 12/31/93 | $20000.00<br>$5000.00 |
| Innocenti , Mario<br>Aerospace Engineering<br>Auburn University, Auburn, AL | PhD | WL/MN | 01/01/93 02/28/94 | $20000.00<br>$12536.00 |
| Jean , Jack<br>Computer Science & Engineering<br>Wright State University, Dayton, OH | PhD | WL/AA | 01/01/93 12/31/93 | $20000.00<br>$34036.00 |
| Jouny , Ismail<br>Electrical Engineering<br>Lafayette College, Easton, PA | PhD | WL/AA | 01/01/93 12/31/93 | $19381.00<br>$4500.00 |
| Kaikhah , Khosrow<br>Computer Science<br>Southwest Texas State College, San Marcos, TX | PhD | RL/IR | 01/01/93 12/31/93 | $20000.00<br>$0.00 |
| Kaw , Autar<br>Mechanical Engineering<br>University of South Florida, Tampa, FL | PhD | WL/ML | 01/01/93 12/31/93 | $20000.00<br>$22556.00 |
| Kheyfets , Arkady<br>Mathematics<br>North Carolina State University, Raleigh, NC | PhD | PL/LI | 01/01/93 12/31/93 | $20000.00<br>$2500.00 |
| Kitchart , Mark<br>Mechanical Engineering<br>North Carolina A & T State University, Greensboro, NC | M.S. | AL/EQ | 01/01/93 12/31/93 | $20000.00<br>$0.00 |
| Koblasz , Arthur<br>Civil Engineering<br>Georgia Institute of Technology, Atlanta, GA | PhD | AL/AO | 01/01/93 12/31/93 | $19826.00<br>$0.00 |
| Koivo , A.<br>Electrical Engineering<br>Purdue University, West Lafayette, IN | PhD | AL/CF | 01/01/93 06/30/94 | $20000.00<br>$0.00 |
| Kundich , Robert<br>Biomedical Engineering<br>University of Tennessee, Memphis, TN | PhD | AL/CF | 01/01/93 12/31/94 | $20000.00<br>$23045.00 |
| Kuo , Spencer<br>Electrical Engineering<br>Polytechnic University, Farmingdale, NY | PhD | PL/GP | 01/01/93 04/30/94 | $20000.00<br>$9731.00 |

# 1993 SREP SUB-CONTRACT DATA

| Report Author<br>Author's University | Author's Degree | Sponsoring Lab | Performance Period | Contract Amount<br>Univ. Cost Share |
|---|---|---|---|---|
| Liou , Juin<br>Electrical and Computer Engineering<br>Universtiy of Central Florida, Orlando, FL | PhD | WL/EL | 01/01/93 12/31/93 | $20000.00<br>$9073.00 |
| Liou , Shy-Shenq<br>Engineering<br>San Francisco State Univesity, San Francisco, CA | PhD | WL/PO | 01/01/93 12/31/93 | $20000.00<br>$13387.00 |
| Manoranjan , Valipuram<br>Pure and Applied Mathematics<br>Washington State University, Pullman, WA | PhD | AL/EQ | 01/01/93 12/31/93 | $19956.00<br>$10041.00 |
| Marks , Dallas<br>Electrical and Computer Engineering<br>University of Cincinnati M.L., Cincinnati, OH | M.S. | WL/AA | 10/01/92 06/30/93 | $20000.00<br>$4731.00 |
| Monsay , Evelyn<br>Physics<br>Le Moyne College, Syracuse, NY | PhD | RL/OC | 01/01/93 12/31/93 | $19634.00<br>$1510.00 |
| Moor , William<br>Industrial & Management Engineering<br>Arizona State University, Tempe, AZ | PhD | AL/HR | 01/01/93 12/31/93 | $20000.00<br>$4833.00 |
| Moore , Carlyle<br>Physics<br>Morehousc College, Atlanta, GA | PhD | AEDC/ | 01/01/93 12/31/93 | $20000.00<br>$4880.00 |
| Mulligan , B.<br>Psychology<br>University of Georgia Research, Athens, GA | PhD | AL/OE | 01/01/93 04/15/94 | $19998.00<br>$13936.00 |
| Murphy , Richard<br>Physics<br>University of Missouri, Kansas City, MO | PhD | PL/WS | 01/01/93 12/31/93 | $20000.00<br>$13022.00 |
| Nilan , Michael<br>Information Studies<br>Syracuse University, Syracuse, NY | PhD | RL/C3 | 01/01/93 12/31/93 | $19998.00<br>$13016.00 |
| Parrish , Allen<br>Computer Science<br>University of Alabama, Tuscaloosa, AL | PhD | RL/C3 | 01/01/93 12/31/93 | $19919.00<br>$20599.00 |
| Piersma , Bernard<br>Chemistry<br>Houghton College, Houghton, NY | PhD | FJSRL/ | 01/01/93 12/31/93 | $20000.00<br>$4000.00 |
| Potasek , Mary<br>Applied Physics<br>Columbia University, New York, NY | PhD | WL/ML | 12/01/93 11/30/93 | $20000.00<br>$7806.00 |
| Qazi , Salahuddin<br>Optical Communications<br>SUNY/Institute of Technology, Utica, NY | PhD | RL/OC | 01/01/93 12/31/93 | $20000.00<br>$68000.00 |
| Reardon , Kenneth<br>Agricultural and Chemical Engineering<br>Colorado State University, Fort Collins, CO | PhD | AL/EQ | 01/01/93 01/31/94 | $19996.00<br>$12561.00 |

# 1993 SREP SUB-CONTRACT DATA

| Report Author<br>Author's University | Author's Degree | Sponsoring<br>Lab | Performance Period | Contract Amoun<br>Univ. Cost Shar |
|---|---|---|---|---|
| Reynolds , David<br>Biomedical & Human Factors<br>Wright State University, Dayton, OH | PhD | AL/CF | 01/01/93 06/30/94 | $20000.00<br>$14063.00 |
| Robinson , Donald<br>Chemistry<br>Xavier University of Louisiana, New Orleans, LA | PhD | AL/OE | 01/01/93 06/30/94 | $20000.00<br>$12935.00 |
| Rodriguez , Armando<br>Electrical Engineering<br>Arizona State University, Tempe, AZ | PhD | WL/MN | 01/01/93 12/31/93 | $20000.00<br>$0.00 |
| Roe , Larry<br>Mechanical Engineering<br>Virginia Polytechnic Inst & State Coll., Blacksburg, VA | PhD | WL/PO | 01/01/93 12/31/93 | $20000.00<br>$11421.00 |
| Romeu , Jorge<br>Assistant Prof. of Mathematics<br>SUNY College at Cortland, Cortland, NY | PhD | RL/OC | 01/01/93 12/31/93 | $19997.00<br>$7129.00 |
| Roppel , Thaddeus<br>Electrical Engineering<br>Auburn University, Auburn, AL | PhD | WL/MN | 01/01/93 12/31/93 | $20000.00<br>$21133.00 |
| Roznowski , Mary<br>Psychology<br>Ohio State University, Columbus, OH | PhD | AL/HR | 01/01/93 03/31/94 | $19953.00<br>$6086.00 |
| Rudzinski , Walter<br>Chemistry<br>Southwest Texas State University, San Marcos, TX | PhD | AL/OE | 01/01/93 12/31/93 | $20000.00<br>$10120.00 |
| Sargent , Robert<br>Engineering and Computer Science<br>Syracuse University, Syracuse, NY | PhD | RL/XP | 01/01/93 12/31/93 | $20000.00<br>$11931.00 |
| Schonberg , William<br>Civil and Environmental Engineering<br>University of Alabama, Huntsville, AL | PhD | WL/MN | 01/01/93 12/31/93 | $19991.00<br>$5083.00 |
| Shaw , Arnab<br>Electrical Engineering<br>Wright State University, Dayton, OH | PhD | WL/AA | 01/01/93 12/31/93 | $20000.00<br>$4766.00 |
| Shively , Jon<br>Engineering & Computer Science<br>California State University, Northridge, CA | PhD | PL/VT | 01/01/93 12/31/93 | $20000.00<br>$9782.00 |
| Slater , Robert<br>Mechanical & Industrial Engineering<br>University of Cincinnati, Cincinnati, OH | M.S. | WL/FI | 01/01/93 12/31/93 | $20000.00<br>$8257.00 |
| Stenzel , Johanna<br>Arts & Sciences<br>University of Houston, Victoria, TX | PhD | PL/LI | 01/01/93 12/31/93 | $20000.00<br>$9056.00 |
| Tan , Arjun<br>Physics<br>Alabama A & M University, Normal, AL | PhD | PL/WS | 01/01/93 12/31/93 | $20000.00<br>$1000.00 |

# 1993 SREP SUB-CONTRACT DATA

| Report Author<br>Author's University | Author's Degree | Sponsoring<br>Lab | Performance Period | | Contract Amount<br>Univ. Cost Share |
|---|---|---|---|---|---|
| Tetrick , Lois<br>Industrial Relations Prog<br>Wayne State University, Detroit, MI | PhD | AL/HR | 01/01/93 | 12/31/93 | $20000.00<br>$17872.00 |
| Tew , Jeffery<br>Industrial & Systems Engineering<br>Virginia Polytechnic Institute, Blacksburg, VA | PhD | RL/IR | 05/31/93 | 12/31/93 | $16489.00<br>$4546.00 |
| Tribikram , Kundu<br>Civil Engineering and Engineering<br>Universtiy of Arizona, Tucson, AZ | PhD | WL/ML | 01/01/93 | 12/31/93 | $20000.00<br>$9685.00 |
| Tuthill , Theresa<br>Electrical Engineering<br>University of Dayton, Dayton, OH | PhD | WL/ML | 01/01/93 | 12/31/93 | $20000.00<br>$24002.00 |
| Venkatasubraman , Ramasubrama<br>Electrical and Computer Engineering<br>University of Nevada, Las Vegas, NV | PhD | WL/ML | 01/01/93 | 12/31/93 | $20000.00<br>$18776.00 |
| Wang , Xingwu<br>Electrical Engineering<br>Alfred University, Alfred, NY | PhD | AL/EQ | 01/01/93 | 12/31/93 | $20000.00<br>$10000.00 |
| Whitefield , Philip<br>Physics<br>University of Missouri, Rolla, MO | PhD | PL/LI | 01/01/93 | 03/01/94 | $20000.00<br>$11040.00 |
| Wightman , Colin<br>Electrical Engineering<br>New Mexico Institute of Mining, Socorro, NM | PhD | RL/IR | 01/01/93 | 12/31/93 | $20000.00<br>$1850.00 |
| Womack , Michael<br>Natural Science and Mathematics<br>Macon College, Macon, GA | PhD | AL/OE | 01/01/93 | 06/30/94 | $19028.00<br>$6066.00 |
| Yuvarajan , Subbaraya<br>Electrical Engineering<br>North Dakota State University, Fargo, ND | PhD | WL/PO | 01/01/93 | 12/31/93 | $19985.00<br>$22974.00 |

# APPENDIX 1:

# SAMPLE SREP SUBCONTRACT

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
1993 SUMMER RESEARCH EXTENSION PROGRAM SUBCONTRACT 93-133

BETWEEN

Research & Development Laboratories
5800 Uplander Way
Culver City, CA 90230-6608

AND

San Francisco State University
University Comptroller
San Francisco, CA 94132

REFERENCE: Summer Research Extension Program Proposal 93-133
Start Date: 01/01/93 End Date: 12/31/93
Proposal Amount: $20,000.00

(1) PRINCIPAL INVESTIGATOR: Dr. Shy Shenq P. Liou
Engineering
San Francisco State University
San Francisco, CA 94132

(2) UNITED STATES AFOSR CONTRACT NUMBER: F49620-90-C-09076

(3) CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER (CFDA): 12.800
PROJECT TITLE: AIR FORCE DEFENSE RESEARCH SOURCES PROGRAM

(4) ATTACHMENTS 1 AND 2: SREP REPORT INSTRUCTIONS

*** SIGN SREP SUBCONTRACT AND RETURN TO RDL***

1. BACKGROUND: Research & Development Laboratories (RDL) is under contract (F49620-90-C-0076) to the United States Air Force to administer the Summer Research Programs (SRP), sponsored by the Air Force Office of Scientific Research (AFOSR), Bolling Air Force Base, D.C. Under the SRP, a selected number of college faculty members and graduate students spend part of the summer conducting research in Air Force laboratories. After completion of the summer tour participants may submit, through their home institutions, proposals for follow-on research. The follow-on research is known as the Summer Research Extension Program (SREP). Approximately 75 SREP proposals annually will be selected by the Air Force for funding of up to $20,000; shared funding by the academic institution is encouraged. SREP efforts selected for funding are administered by RDL through subcontracts with the institutions. This subcontract represents such an agreement between RDL and the institution designated in Section 5 below.

2. RDL PAYMENTS: RDL will provide the following payments to SREP institutions:

- 90 percent of the negotiated SREP dollar amount at the start of the SREP Research period.
- the remainder of the funds within 30 days after receipt at RDL of the acceptable written final report for the SREP research.

3. INSTITUTION'S RESPONSIBILITIES: As a subcontractor to RDL, the institution

designated on the title page will:

a. Assure that the research performed and the resources utilized adhere to those defined in the SREP proposal.

b. Provide the level and amounts of institutional support specified in the RIP proposal.

c. Notify RDL as soon as possible, but not later than 30 days, of any changes in 3a or 3b above, or any change to the assignment or amount of participation of the Principal Investigator designated on the title page.

d. Assure that the research is completed and the final report is delivered to RDL not later than twelve months from the effective date of this subcontract, but no later than December 31, 1993. The effective date of the subcontract is one week after the date that the institution's contracting representative signs this subcontract, but no later than January 15, 1993.

e. Assure that the final report is submitted in accordance with Attachment 1.

f. Agree that any release of information relating to this subcontract (news releases, articles, manuscripts, brochures, advertisements, still and motion pictures, speeches, trade association meetings, symposia, etc.) will include a statement that the project or effort depicted was or is sponsored by: Air Force Office of Scientific Research, Bolling AFB, D.C.

g. Notify RDL of inventions or patents claimed as the result of this research as specified in Attachment 1.

h. RDL is required by the prime contract to flow down patent rights and technical data requirements in this subcontract. Attachment 2 to this subcontract contains a list of contract clauses incorporated by reference in the prime contract.

4. All notices to RDL shall be addressed to:

RDL Summer Research Program Office
5800 Uplander Way
Culver City, CA 90230-6608

5. By their signatures below, the parties agree to the provisions of this subcontract.

______________________________
Abe S. Sopher
RDL Contracts Manager

______________________________
Signature of Institution Contracting Official

______________________________
Typed/Printed Name

______________________________
Date

______________________________
Title

______________________________
Institution

______________________________(Date/Phone)

## ATTACHMENT 2
## CONTRACT CLAUSES

This contract incorporates by reference the following clauses of the Federal Acquisition Regulations (FAR), with the same force and effect as if they were given in full text. Upon request, the Contracting Officer or RDL will make their full text available (FAR 52.252-2).

| FAR CLAUSES | TITLE AND DATE |
|---|---|
| 52.202-1 | DEFINITIONS (SEP 1991) |
| 52.203-1 | OFFICIALS NOT TO BENEFIT (APR 1984) |
| 52.203-3 | GRATUITIES (APR 1984) |
| 52.203-5 | COVENANT AGAINST CONTINGENT FEES (APR 1984) |
| 52.304-6 | RESTRICTIONS ON SUBCONTRACTOR SALES TO THE GOVERNMENT (JUL 1985) |
| 52.203-7 | ANTI-KICKBACK PROCEDURES (OCT 1988) |
| 52.203-12 | LIMITATION ON PAYMENTS TO INFLUENCE CERTAIN FEDERAL TRANSACTIONS (JAN 1990) |
| 52.204-2 | SECURITY REQUIREMENTS (APR 1984) |
| 52.209-6 | PROTECTING THE GOVERNMENT'S INTEREST WHEN SUBCONTRACTING WITH CONTRACTORS DEBARRED, SUSPENDED, OR PROPOSED FOR DEBARMENT (NOV 1992) |
| 52.212-8 | DEFENSE PRIORITY AND ALLOCATION REQUIREMENTS (SEP 1990) |
| 52.215-1 | EXAMINATION OF RECORDS BY COMPTROLLER GENERAL (APR 1984) |
| 52.215-2 | AUDIT - NEGOTIATION (DEC 1989) |
| 52.222-26 | EQUAL OPPORTUNITY (APR 1984) |
| 52.222-28 | EQUAL OPPORTUNITY PREAWARD CLEARANCE OF SUBCONTRACTS (APR 1984) |

| | |
|---|---|
| 52.222-35 | AFFIRMATIVE ACTION FOR SPECIAL DISABLED AND VIETNAM ERA VETERANS (APR 1984) |
| 52.222-36 | AFFIRMATIVE ACTION FOR HANDICAPPED WORKERS (APR 1984) |
| 52.222-37 | EMPLOYMENT REPORTS ON SPECIAL DISABLED VETERAN AND VETERANS OF THE VIETNAM ERA (JAN 1988) |
| 52.223-2 | CLEAN AIR AND WATER (APR 1984) |
| 52.232-6 | DRUG-FREE WORKPLACE (JUL 1990) |
| 52.224-1 | PRIVACY ACT NOTIFICATION (APR 1984) |
| 52.224-2 | PRIVACY ACT (APR 1984) |
| 52.225-13 | RESTRICTIONS ON CONTRACTING WITH SANCTIONED PERSONS (MAY 1989) |
| 52.227-1 | AUTHORIZATION AND CONSENT (APR 1984) |
| 52.227-2 | NOTICE AND ASSISTANCE REGARDING PATENT AND COPYRIGHT INFRINGEMENT (APR 1984) |
| 52.227-10 | FILING OF PATENT APPLICATIONS - CLASSIFIED SUBJECT MATTER (APR 1984) |
| 52.227-11 | PATENT RIGHTS - RETENTION BY THE CONTRACTOR (SHORT FORM) (JUN 1989) |
| 52.228-6 | INSURANCE - IMMUNITY FROM TORT LIABILITY (APR 1984) |
| 52.228-7 | INSURANCE - LIABILITY TO THIRD PERSONS (APR 1984) |
| 52.230-5 | DISCLOSURE AND CONSISTENCY OF COST ACCOUNTING PRACTICES (AUG 1992) |
| 52.232-23 | ASSIGNMENT OF CLAIMS (JAN 1986) |
| 52.237-3 | CONTINUITY OF SERVICES (JAN 1991) |

|  |  |
|---|---|
| 52.246-25 | LIMITATION OF LIABILITY - SERVICES (APR 1984) |
| 52.249-6 | TERMINATION (COST-REIMBURSEMENT) (MAY 1986) |
| 52.249-14 | EXCUSABLE DELAYS (APR 1984) |
| 52.251-1 | GOVERNMENT SUPPLY SOURCES (APR 1984) |

# APPENDIX 2:

# SAMPLE TECHNICAL EVALUATION FORM

# 1993 SUMMER RESEARCH EXTENSION PROGRAM

**RIP NO.: 93-0092**
**RIP ASSOCIATE: Dr. Gary T. Chapman**

Provided are several evaluation statements followed by ratings of (1) through (5). A rating of (1) is the lowest and (5) is the highest. Circle the rating level number you best feel rates the statement. Document additional comments on the back of this evaluation form.

Mail or fax the completed form to :

RDL
Attn: 1993 SREP TECH EVALS
5800 Uplander Way
Culver City, CA 90230-6608
(FAX: 310 216-5940)

1. This SREP report has a high level of technical merit. 1 2 3 4 5
2. The SREP program is important to accomplishing the labs's mission 1 2 3 4 5
3. This SREP report accomplished what the associate's proposal promised. 1 2 3 4 5
4. This SREP report addresses area(s) important to the USAF 1 2 3 4 5
5. The USAF should continue to pursue the research in this SREP report 1 2 3 4 5
6. The USAF should maintain research relationships with this SREP associate 1 2 3 4 5
7. The money spent on this SREP effort was well worth it 1 2 3 4 5
8. This SREP report is well organized and well written 1 2 3 4 5
9. I'll be eager to be a focal point for summer and SREP associates in the future. 1 2 3 4 5
10. The one-year period for complete SREP research is about right 1 2 3 4 5

******USE THE BACK OF THIS FORM FOR ADDITIONAL COMMENTS******

LAB FOCAL POINT'S NAME (PRINT): ______________________________

OFFICE SYMBOL: ____________________ PHONE: ________________

# EVALUATION OF VARIABLE STRUCTURE CONTROL FOR MISSILE AUTOPILOTS USING REACTION JETS AND AERODYNAMIC CONTROL

FINAL REPORT

by
Ajay Thukral, John E. Cochran, Jr.
Department of Aerospace Engineering, Auburn University, Alabama 36849-5338

and
Mario Innocenti
Department of Electrical Systems and Automation
University of Pisa, 56126 Pisa, Italy

submitted to
Research Development Laboratories
5800 Uplander Way, Culver City, California 90230-6608

under
Contract: RDL-93-132

Auburn University, Alabama
15 May 1994

## Preface

This report documents the results obtained under the grant RDL-93-132, from February 1993 until February 1994. The work was performed at Auburn University, Alabama, in the Department of Aerospace Engineering. The principal investigator of record for the project was Dr. John E. Cochran, Jr., however, the work was initiated by Dr. Mario Innocenti while he was at Auburn. Furthermore, most of the work was done by Dr. Mario Innocenti, as a consultant, and Mr. Ajay Thukral, a Ph.D. candidate. Mr. Gregory D. Strawn, graduate student, also contributed to the preparation of the report.

John E. Cochran, Jr.
Principal Investigator

## Table of Contents

Appendices

## List of Figures

List of Figures (contd.)

**List of Figures (contd.)**

## List of Tables

## List of Important Symbols

| | |
|---|---|
| RCS | reaction jet thrusters |
| TVC | thrust vector control |
| VSC, VSS | variable structure control |
| $x$, $y$, $u$, $u_T$ | state, output, input vectors |
| $x_m$, $u_m$ | state, input model vectors |
| A, B, C | system matrices |
| L, M, N, K, $\rho$ | gain matrices |
| $\delta$ | elevator deflection, smoothing parameter for VSS |
| Re | Reynolds number |
| L, D | lift, drag |
| $C_i$ | lift, drag, normal force, pitching moment coefficients as appropriate |
| $\alpha$ | angle of attack |
| V | velocity |
| $q$, $\theta$ | pitch rate, pitch angle |
| $\gamma$ | flight path angle |
| $\Theta$, $\Gamma$, X, $X_m$ | trim values of appropriate variables/vectors |

## 1. Introduction

The feasibility of combining traditional aerodynamic control with reaction jets, in the framework of missile autopilot design, was addressed in this work. The purpose of propulsive actuation is mainly to increase the angle of attack envelope for improved turn rate capabilities and maneuverability. Due to nonlinear characteristics of both controller and airframe dynamics, aerodynamic and geometric model uncertainties, a control strategy based on variable structure systems was adopted. A control law was then synthesized for a simplified pitch channel autopilot and used in a high angle of attack midcourse maneuver. Results of the nonlinear simulation show the capability of the autopilot to satisfy the control objectives for a variety of flight conditions.

### 1.1. Motivation

Future missile systems will be required to possess higher turn rates and larger maneuverability envelopes, while simultaneously meeting the requirement of reduced storage and signature. In this respect, efforts are under way to evaluate alternate methods of missile control as opposed to purely aerodynamic control [1], [2], [3].

Several technology payoffs can be envisioned if alternate control strategies are implemented, among which there are:

- decreased stowage volume for internal carriage, especially important for the type of fighters currently being developed,
- increased maneuverability and off-boresight capability for improved all-aspect defensive shield,
- high angle of attack launch capability to take advantage of improved aircraft agility, and better end-game accuracy.

The achievement of these payoffs poses difficult challenges to the control system designer that encompasses all phases of flight. For example, during separation, an increase in pitch-up tendencies can be expected due to lack of sufficient aerodynamic stabilization to achieve high maneuverability and high angles of attack. In the midcourse phase, the system may be required to perform fast 180-degree turns to account for defense and engagement against tail-positioned threats. During the end-game, the reduced aerodynamic control effectiveness due to limited fin size must be appropriately compensated for in order to generate sufficient load factors in a very short time.

The desire to limit its cross section and volume drastically reduces the amount of aerodynamic effectiveness of a missile. This loss in control power must be compensated for and/or augmented by using alternate technologies. Possible options are reaction-based control in the form of thrust vectoring (TVC) and/or a reaction jet thrusters (RCS). A generic configuration based on three possible control sources is shown in Figure 1.

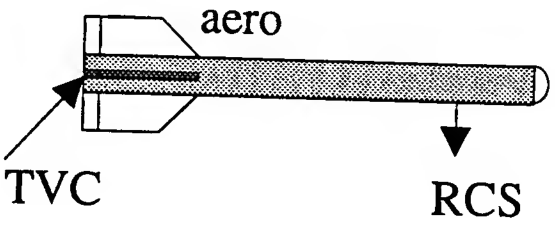


Figure 1. Generic Control Cconfiguration

The potential modifications involving the implementation of propulsion control and its integration with aerodynamic surfaces are several, and their description and implications are beyond the scope of the present research. Just to summarize some of the aspects, however, we mention the technology involved with the design of each component, as well as the integration of elements leading to variable degree of effort: from the mere addition of actuator on existing airframe, all the way to a new missile design. The work done under this grant was concentrated on one of the propulsive solutions, specifically the use of reaction jets. The application of thrust vector control is addressed in reference [4].

## 1.2. Maneuver Description

In order to gain appreciation for some of the problems involving reaction jet control and its blending with traditional aerodynamic control, a high angle of attack midcourse maneuver was chosen as test scenario. In particular, a two-dimensional, heading reversal trajectory in the longitudinal plane was selected as a typical defensive maneuver against tail and fly-by threats as shown schematically in Figure 2.

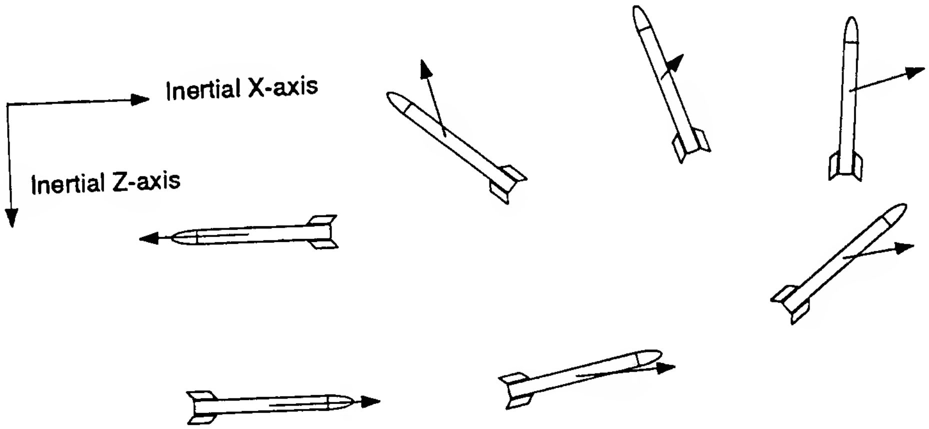


Figure 2. Selected Midcourse Trajectory

Many challenges to guidance and control systems are posed by the above selection. To completely overcome them will require much greater effort than that available during the present research. However, some of the critical issues are addressed here leading to a preliminary design of the autopilot.

The maneuver is a 180-degree off-boresight trajectory with turn rates of the order of 80 deg/sec, capable of pointing as well as flying the missile roughly in the opposite direction as quickly as possible along a minimum radius turn path and in a time frame of the order of two seconds.

The specifications involve both guidance and autopilot requirements. The guidance aspects deal with the generation of an appropriate flight path along which the missile turns in minimum time changing its heading and an attitude of up to 180 degrees. The selection of this path could depend on agility issues and/or tactical ones. The autopilot aspects deal with the creation of forces and moments on the missile capable of generating accelerations and attitude rates required by the guidance system. Appropriate blending of aerodynamic and reaction jet controls may be required since, during parts of the trajectory, the missile may experience loss of lifting capabilities due to angles of attack much higher than stall.

In this report we do not address the question of guidance law design, rather we present the development of a nonlinear autopilot logic capable of implementing the maneuver, and a blending strategy which uses aerodynamic control at low angles of attack and RCS control when the missile angle of attack is higher stall.

### 1.3. Summary of Results

The results provided in this report are in terms of autopilot structure, gains and simulation data. The theory of variable structure control is briefly reviewed first, then a description of the model dynamics derivation at low and high angle of attack is presented to set the analytical framework for the autopilot design.

The design of the autopilot is the central part of the report. The design includes: control structure, controller gain matrices, and block diagrams. The performance of the closed loop system is evaluated using a nonlinear simulation code that contains attitude as well as point mass dynamics of the missile. The simulation code was written using Matlab® and the software is included with this report as part of the deliverables.

## 2. Variable Structure Control

Variable structure control has been described in the former Soviet literature since the early sixties, see, for example, Emel'yanov [5], Utkin [6] and Itkis [7], among others. Invariance of VSC to a class of disturbances and parameter variations was first developed by Drazenovic in 1969 [8]. In the past two decades, a large amount of research has been performed in the area by the international community. This research has linked VSC to adaptive control and model reference adaptive control, using Lyapunov control techniques. Also, investigators have derived connections of VSC with hyperstability theory, and solved VSC tracking problems (see references [9] and [10] for a survey on the subject).

Most of the applications of VSC have been in the areas of industrial control and robotics. Only recently some work has been done in the aerospace field. Applications to aircraft control have been presented by Calise and Kramer [11] where robustness with respect to nonlinearities is addressed, and by Innocenti and Thukral [20]. Mudge and Patton [12], solved the sensitivity to parameter variations by incorporating eigenstructure assignment in the structure of the control law, Hedrick et al. [13] used Slotine's concept of boundary layer to eliminate chattering. Lyapunov stability theory and VSC were used by Vadali in designing large-angle maneuvers controllers for a spacecraft [14]. Applications to missiles appear to have been confined mainly to guidance schemes [15], [16].

The essential feature of a variable structure controller is that it uses nonlinear feedback control with discontinuities on one or more manifolds (sliding hyperplanes) in the state space, or error space, in the case of model following control. This type of methodology is attractive in the design of controls for nonlinear, uncertain, dynamic systems with uncertainties and nonlinearities of unknown structure as long as they are bounded and occurring within a subspace of the state space [9]. Ryan and Corless [17] have also shown that VSC could be used to establish 'almost certain' convergence to vicinity of the origin for a class of uncertain systems. A brief description of the principles of variable structure systems is now presented, and essentially follows those of references [6] and [9].

The basic feature of VSC is sliding motion. This occurs when the system state continuously crosses a switching manifold because all motion in its vicinity is directed towards the sliding surface. When the motion occurs on all the switching surfaces at once, the system is said to be in the "sliding mode" and then the original system is equivalent to an unforced, completely controllable system of lower order.

The design of a variable structure controller consists of several steps: the choice of switching surfaces, the determination of the control law and the switching logic associated with the discontinuity surfaces (usually fixed hyperplanes that pass through the origin of the state space). To ensure that the state reaches the origin along the sliding surfaces, the equivalent system must be asymptotically stable. This requirement defines the selection of the switching hyperplanes (sometimes called the “existence” problem), which is completely independent of the choice of control laws. The selection of the control law is the so-called "reachability" problem. It requires that the system be capable of reaching the sliding hypersurface from any initial state.

During operation in the sliding mode, the discontinuous control chatters about the switching surface at high frequency. Chatter is the major problem associated with this type of control. Execution of control commands may require high energy effort from the actuators, thus leading to continuous saturation. It can also excite neglected high order dynamics. This is perhaps the reason why VSC has not yet found wider acceptance in the flight control community, where smoothness of actuation is desirable to avoid saturation and, possibly, instability. The introduction of discontinuous actuators such as reaction jets and active flow control is however changing this perspective and variable structure systems are being viewed as a viable alternative to traditional relay control strategies.

There are several ways to mitigate the effects of chattering, with little loss in performance. These include the definition of a boundary layer near the sliding surface as introduced by Slotine, and/or the introduction of a smoothing parameter in a unit vector-type control law as shown by Ambrosino et al [18], Burton and Zinober [9], Balestrino [19], and Thukral and Innocenti [20]. The latter approach was used in the present work.

As noted in [21], the smoothing factors do not guarantee full robustness, however such relaxation is the price paid for avoiding actuator saturation. Of course, smoothing is not necessary when on-off actuators such as thrusters are being used.

The general control problem is based on the following nonlinear, uncertain, and controllable dynamic system

$$\dot{x} = (A + \Delta A)x + (B + \Delta B)u + Cv \tag{1}$$

$$y = x + w$$

where the state and input vectors have dimensions n and m respectively, *v(t)* is a one-dimensional disturbance vector also representing nonlinearities and *w(t)* is a vector of output (measurement) uncertainties. The parameter variation matrices $\Delta A$ and $\Delta B$ can be uncertain and time varying. Matching conditions are assumed to be satisfied by the matrices $\Delta A$, $\Delta B$ and $C$, thus satisfying Drazenovic invariance conditions as well as perfect model following [8]. Since matching requires $\Delta A$, $\Delta B$ and $C$ to be in the range space of $B$ (assumed to be full rank), the following relations are necessary for perfect invariance

$$\begin{aligned} \Delta A &= BD, \\ \Delta B &= BE, \\ C &= BF \end{aligned} \tag{2}$$

where $D$, $E$, and $F$ have dimension nxn, mxm and mxl respectively. The purpose of a VSC design is then to determine the control law $u$ of the form

$$u_i(x) = \begin{cases} u_i^+ \; for \, s_i(x) > 0 \\ u_i^- \; for \, s_i(x) < 0 \end{cases} \tag{3}$$

with the switching hyperplanes denoted in matrix form by

$$s = Gx \tag{4}$$

where $s$ is m-dimensional and $G$ is an mxn constant matrix. For a stable sliding motion to occur on all surfaces, the following conditions, based on Lyapunov's stability theory, must be satisfied:

$s_i \dot{s}_i < 0$ near $s_i = 0$ $\quad s = Gx = \dot{s} = G\dot{x} = O$ in the sliding mode. (5)

Since the sliding mode belongs to the null space of $G$, if the product $GB$ is nonsingular, the sliding motion is independent of the control law. During sliding, from Eqs. (1) and (5) we can determine an equivalent control law

$$\begin{aligned} u_{eq} &= -\,(GB)^{-1}\, G\, [Ax + h\,] \\ h &= \Delta Ax + \Delta Bu + Cv \end{aligned} \tag{6}$$

Since the matching conditions (2) are assumed to be valid, the system dynamics during sliding are then governed by

$$\dot{x} = \left[I - B(GB)^{-1}G\right]Ax \tag{7}$$

*showing the sliding motion to be insensitive to unknown, but bounded, parameter variations and disturbances*. The selection of the switching surfaces, i. e. *G*, depends on the desired system behavior during sliding and given by Eq. (7).

To select the switching surfaces, we consider first a nominal system extracted from (1) and given by

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= x \\ s &= Gx \end{aligned} \tag{8}$$

In order to simplify the design scheme, we transform Eq. (8) into a controllable canonical form using the transformation $q = Tx$, where *T* is an orthogonal matrix. This yields, [with $A_{11}$ square of dimensions (n-m)]

$$\begin{aligned} \dot{q} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} q + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} u \\ y &= T^T q. \\ s &= GT^T q = \begin{bmatrix} G_1 & G_2 \end{bmatrix} q \end{aligned} \tag{9}$$

Note that, since *GB* is nonsingular, so are $G_2B_2$ and $G_2$. During sliding, we have from $s = O$

$$\begin{aligned} \dot{q}_1 &= [A_{11} - A_{12}K]q_1 \\ q_2 &= -Kq_1 \end{aligned} \tag{10}$$

with $K = G_2^{-1}G_1$.

The sliding motion occurs, therefore, in the n-m dimensional subspace of the state space. The choice of *K*, and consequently of *G*, is free for the designer to choose and several methods have been used in the literature such as pole placement, eigenstructure assignment [12], and optimal control [20]. Using the latter method to find $K$, we can set up an LQR synthesis that minimizes

$$J=\frac{1}{2}\int_{t}^{\infty}\left[x^T Qx\right]\,dt \quad \text{with } Q>0$$

subject to the constraints given by Eq. (10). The above index of performance can be reduced to the transformed state space $q$ by using $T$. We can write

$$TQT^T=\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

If we let

$$\begin{cases} Q^*=Q_{11}-Q_{12}Q_{22}^{-1}Q_{21} \\ A^*=A_{11}-A_{12}Q_{22}^{-1}Q_{21} \\ \varsigma=q_2+Q_{22}^{-1}Q_{21}q_1 \end{cases} \tag{11}$$

then the LQR problem has now the standard form

$$\begin{cases} J=\frac{1}{2}\int_{t}^{\infty}\left[q_1^T Q^* q_1+\varsigma^T Q_{22}\varsigma\right]dt \\ \dot{q}_1=A^* q_1+A_{12}\varsigma. \end{cases} \tag{12}$$

After solving for the appropriate Riccati matrix $P$ associated with Eq. ( 12), we obtain

$$K=Q_{22}^{-1}\left[Q_{21}+A_{12}^T P\right]. \tag{13}$$

A simple method for deriving the switching matrix $G$ in (10) from Eq. (13) is given in [9] and [20]. If we let $G_2=I_m$, then $\begin{bmatrix} G_1 & G_2 \end{bmatrix}=GT^T=\begin{bmatrix} K & I_m \end{bmatrix}$, thus

$$G=\begin{bmatrix} K & I_m \end{bmatrix}T. \tag{14}$$

Having specified the sliding surfaces, we now turn our attention to the computation of the control law $u$, that will drive the state vector $x$ into the null space of

$G$ and maintain it there. The choice of control is only limited by the discontinuity on one or more subspaces containing the null space of $G$ as stated in Eq. (3).

In general, the VSC control law $u$ consists of a linear component $u^L$ and a nonlinear one $u^N$, combined together to produce the feedback, with the nonlinear component incorporating the discontinuous elements. In the present work, the following initial structure for the control law is chosen to be

$$u = u^L + u^N = Lx + \rho \frac{Nx}{\|Mx\|}. \tag{15}$$

The linear component is typically a full state feedback, while the nonlinear element has a unit vector form [9], [19] that is easier to implement than other structures. The parameter matrix ρ is free to be chosen and the matrices $N$, $M$, and $G$ [see Eq. (14)] belong to the same null space.

To compute the gain matrices $L$, $M$, and $N$ in Eq. (15), we follow the procedure described in [17] and [20], or we use a simple sign function depending on the phase of flight as described in section 4.

Let us define a new nonsingular transformation matrix $T_2$ as

$$T_2 = \begin{bmatrix} I_{n-m} & 0 \\ K & I_m \end{bmatrix}$$

Using the above matrix the state vector $q$ is changed into $z = T_2\, q$, with $z_1 = q_1$ and $z_2 = Kq_1 + q_2$. The dynamics of $z$ are then given by

$$\begin{cases} \dot{z}_1 = \Lambda_1 z_1 + A_{12} z_2 \\ \dot{z}_2 = \Lambda_2 z_1 + \Lambda_3 z_2 + B_2 u \end{cases} \tag{16}$$

where

$$\begin{cases} \Lambda_1 = A_{11} - A_{12}K \\ \Lambda_2 = K\Lambda_1 + A_{21} - A_{22}K\,. \\ \Lambda_3 = KA_{12} + A_{22} \end{cases} \tag{17}$$

To attain a sliding mode it is required from (10) and (16) that $z = \dot{z} = 0$, therefore we can define

$u(z) = u^L + u^N$ where

$$u^L(z) = -B_2^{-1}\left[\Lambda_2 \quad \left(\Lambda_3 - \Lambda_3^*\right)\right]z = -\Theta\ z \tag{18}$$

where $\Lambda_3^*$ is a stability matrix whose eigenvalues determine the speed and transient characteristics with which the state vector *asymptotically* attains a sliding motion. The nonlinear component allows the state $z_2$ to reach the sliding mode in *finite time.* By defining $P_1 > 0$ to be the solution of the Lyapunov equation

$$P_1\Lambda_3^* + (\Lambda_3^*)^T P_1 + I_m = 0$$

we can set

$$u^N = -\rho\frac{B_2^{-1}P_1 z_2}{\|P_1 z_2\|}. \tag{19}$$

Finally, returning to the original state vector $x$ we have the control law given by Eq. (15), with gain matrices

$$\begin{cases} L = -\Theta\ T_2 T \\ N = -B_2^{-1}\left[0 \quad P_1\right]T_2 T. \\ M = \left[0 \quad P_1\right]T_2 T \end{cases} \tag{20}$$

When there are disturbances and parameter variations included in the system dynamics as in Eq. (1), the control law (15) and gain matrices (20) still hold, the output vector $y$ however appears in the control structure in place of $x$ and $\rho$ becomes a function of the off-nominal components $\Delta A$, $\Delta B$, etcetera. Details on the computation of $\rho$ can be found in [10] and [17]. Briefly, recalling the uncertain system model (1), and using Eqs. (2) and (15), we obtain the form

3. Select the speed with which sliding is to be attained by choosing $\Lambda_3^*$ in Eq. ( 18)

4. Compute the control gain matrices using Eq. (20)

5. Select ρ according to the perturbations included in the model, else choose it to be a constant

6. Implement a smoothed control law by proper choice of δ in Eq. (23)

The general procedure described above will be specialized and applied to the autopilot design in section 4.

## 3. Missile Dynamics

This section describes the derivation of governing equations of motion for the system's dynamics and reviews the underlying theory behind the modelling of the aerodynamic characteristics.

### 3.1. System Characteristics

Before we can design an autopilot, a model of the system to be controlled must be available. Since the flight envelope of interest here includes both low and high angle of attack conditions, two dynamic models were established, the first is based on the standard short period mode approximation. The second is a combination of pure pitching motion and point mass dynamics. From the viewpoint of aerodynamics, the system dynamic models are based on a generic air-to-air configuration corresponding to a standard cruciform axial-symmetric shape shown in Figure 3. Preliminary analyses [2], [4], indicated that structure flexibility was not a crucial issue for such geometry. The estimated first bending mode natural frequency is of the order of 30 Hertz and outside the projected autopilot bandwidth. For this reason, the system was modelled as a rigid body. That is no bending dynamics. which may require filtering were included in the present work. The rigid body hypothesis, however, needs to be addressed in a future follow-on activity where different length to caliber ratios are investigated. The main geometric characteristics are listed in Table 1.

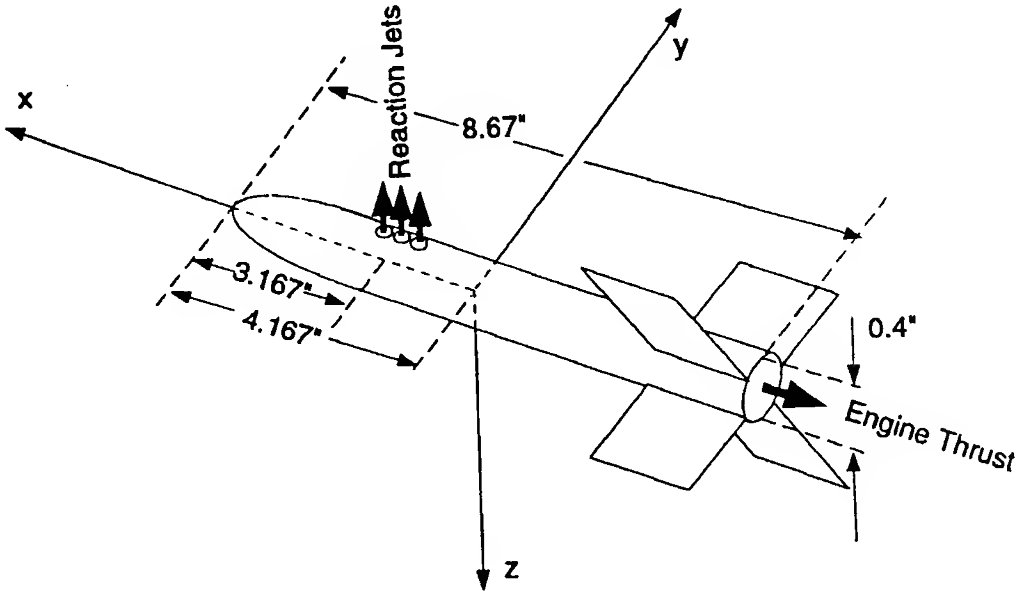


Figure 3. Missile Configuration

Table 1. Physical and Geometric Characteristics

| | |
|---|---|
| $L_{REF}$ | 0.4167 ft |
| S | 0.1367 sqft |
| mass | 7.0 slugs |
| $I_y = I_z$ | 51.0 sl-sqft |
| $I_x$ | 0.229 sl-sqft |
| Fins | X configuration |
| $L_{RCS}$ | 3.167 ft from tip |
| xcg | 4.167 ft from tip |
| Length | 8.67 ft |
| Diameter | 0.4 ft |

The aerodynamic control forces and moments are generated by deflecting fins. The fins are smaller than traditional ones. The sign convention is taken from [22], which defines a positive panel (surface) deflection as one that produces a negative rolling moment increment at zero angle of attack and sideslip. This sign convention, along with the relative panel deflections for pitch, roll, and yaw control, respectively, are illustrated in Figure 4 and Table 2.

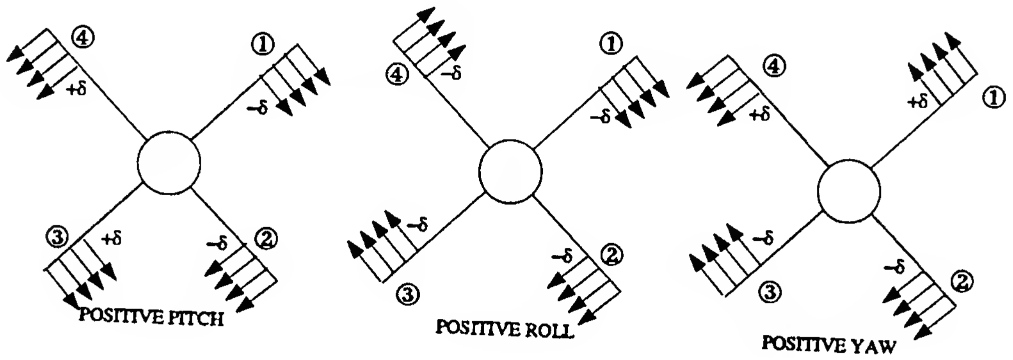


Figure 4. Control Panel Deflections (from rear)

Table 2. Moments and Panel Deflections

| Moment | Pan-1 | Pan-2 | Pan-3 | Pan-4 |
|---|---|---|---|---|
| Pitching | –δ | –δ | +δ | +δ |
| Rolling | –δ | –δ | –δ | –δ |
| Yawing | +δ | –δ | –δ | +δ |
| Neutral | +δ | –δ | +δ | –δ |

By convention, pitch-up corresponds to a positive pitching moment. For a flight vehicle, this corresponds to a negative "elevator" deflection. If panel 1 is selected as reference, the sign convention is the standard one. Similarly for roll and yaw rotations. The fin actuators were modelled as linear first order systems in the present analysis.

The propulsion system consists of a set of reaction jets (RCS) and a main engine. Because this work is preliminary in nature, the location, size, and detailed operational characteristics of the thrusters are not discussed here in any detail. The interaction between aerodynamic flow and jet plumes has also been neglected up to this point.

For the purpose of the present study, the actuation characteristics of the thrusters were modelled as those of a typical relay, with a constant output thrust, chosen nominally as 500 lbs, and a first order lag as shown schematically in Figure 5. During the simulation, a parametric analysis was carried out using different values of nominal thrust.

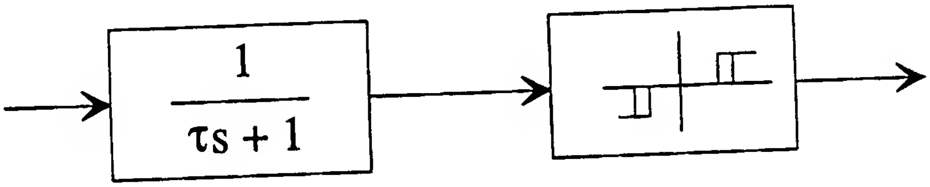


Figure 5. RCS Model

The main engine, which in principle could have thrust vectoring capabilities, was assumed operating at a nominal thrust $T_E$ = 5,000 lbs. The engine was used during the post-stall phase of the maneuver when boosting was needed in order to recover the dynamic pressure lost and to provide velocity vector rotation. The firing time interval during this phase was another parameter varied in the simulation analysis.

The nominal flight condition was chosen to be that of Mach 0.8 and altitude of 10,000 feet. A summary of flight condition and propulsion data is given in Table 3.

Table 3. Missile and Flight Condition Data

| | |
|---|---|
| Main Engine Nominal | $T_E$ = 5,000 lbs |
| Reaction Jets Nominal | $T_{RCS}$ = 500 lbs |
| RCS Time Constant | $\tau_u$ = 1/500 sec |
| Elevator Time Constant | $\tau_\delta$ = 1/180 sec |
| RCS Deadband | variable by design |
| Reference Mach Number | M = 0.8 |
| Trim Altitude | h = 10,000 ft |
| Trim Angle of Attack | 10 degrees |
| Trim Attitude | 10 degrees |

## 3.2. Missile Aerodynamics

The aerodynamic forces and moments are usually obtained from wind tunnel data of the vehicle and then "tuned" using flight testing. In the present work, neither type data was available and analytical and numerical prediction methods were used. The uncertainty and parameter variations introduced this way were then used as robustness test for the variable structure controller.

In the maneuver chosen as the test scenario for autopilot validation, the vehicle experiences a wide range of variations in angle of attack . Due to the absence of data, the computation of aerodynamic coefficients was carried out by considering two different flight regimes. First, the missile DATCOM code [22] was used for low angles of attack (predefined by being below an assumed stall value between 35 and 40 degrees). Second, classical fluid dynamics prediction methods [23], [24] were used for high angles of attack (up to 90 degrees).

For bluff bodies, including streamline bodies at high angles of attack, the flow separates causing a large wake behind the body. The predominant component of the drag force is therefore pressure drag. The estimation of the aerodynamic forces was done assuming the missile as a cylinder and neglecting the interference effect between wings and main body as a first approximation. For the present work it was also assumed that the center of pressure was coincident with the geometric center of the missile.

The main aerodynamic force at high angles of attack is the normal force $N$. The normal force coefficient $C_N = N/QS$ , where $Q$ is the dynamic pressure and $S$ the reference area, is a function of (1) angle of attack, (2) Reynolds number, and, (3) Mach number. The coefficient $C_N$ was first computed as a function of Re, at zero angle of attack and constant Mach number, then modified accordingly. Based on the above assumptions, a code was written to obtain aerodata for high angle of attack values.

Reynolds Number (Re)

For very low Re, the flow, described as " creeping flow ", creates pressure differentials equivalent to skin friction. For symmetric, elliptical cylinders, the drag coefficient based on the frontal area is given by

$$C_{D0} = 8\pi / R_d / \left[ c/(c+h) + 1.5 - 2.3\ln(R_*) \right] \tag{24}$$

where h is the height and c is the length of the axis of the cylinder in the flow direction as shown in Figure 6. The Reynolds number $R_*$ is defined as

$$R_* = V(h+c)/2/\upsilon \tag{25}$$

For a cylinder, $R_d = Vd/\upsilon$, since $h = c$ and $d = (h + c)/2$. Thus, for a cylinder with $R_d < 1$, the drag coefficient at zero angle of attack is

$$C_{D0} = 10.9 / R_d / (0.87 - \ln R_d) \tag{26}$$

If $R_d > 1$, dynamic forces of the fluid cannot be neglected since they influence the $C_{D0}$. Dynamic forces are predominant over the viscous forces to such an extent that they cause periodic shedding of vortices behind blunt bodies at a non-dimensional frequency, which increases steadily with the Reynolds number. This type of vortex pattern is found in the wake of 2-D bodies, such as cylinders, plates or bluff rods.

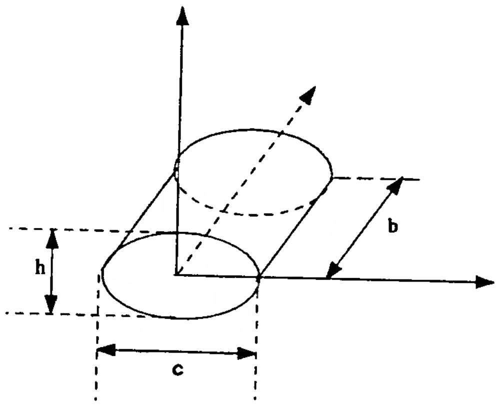


Figure 6. Elliptical Cylinder and Dimensions

This comparatively stable system is called "double-row vortex trail" or "vortex-street". Straight vortices are periodically released from the two sides of the body. If the cylinder is moving with respect to fluid at a velocity w, the vortex street follows the cylinder at velocity w. The distance between two vortices is x, as shown in Figure 7. Both w and x are difficult to predict theoretically. Below the critical Reynolds number, $w/V = 1/6$ and $x/d \sim 4.5$(based on tests), and $C_{D0} = 4.5C_{Dx}$, where

$$C_{Dx} = D/(qbx) = 1.6(w/V) - 0.64(w/V)^2$$

where, b is the span of the cylinder. It has been shown that the drag equivalent for a bluff body, such as a cylinder is entirely contained in the vortex system. It is emphasized that the vortex street is a mechanism which leads to a realistic drag coefficient, without introducing any qualitative viscosity values. The frequency indicating the "Strouhal number," reaches a constant level in the vicinity of $R_d = 10^3$. For regions from 1 to $10^3$ a curve fit is a good enough approximation. Accurate calculation of $C_{D0}$ requires however the knowledge of pressure coefficient, $C_p$.

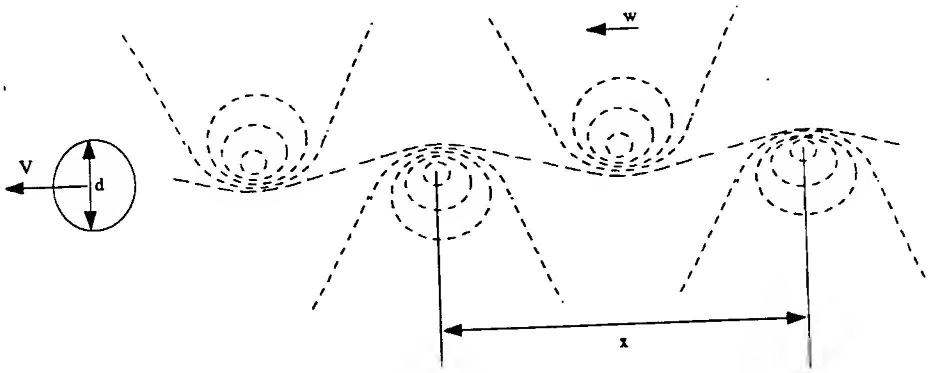


Figure 7. Vortex Street

Vortex frequency

The number of vortices formed at one side of the street in a unit time is given by

$$f = (V-w)/x \tag{27}$$

where $w = V/6$, and $x = 4.5\ d$. The Strouhal number $S$, is defined as

$$S = \textit{Strouhal number} = f\,h\,/V \tag{28}$$

where $h = d$ for cylinder. For a flat plate, h is the height of the plate. The drag coefficient $C_{D0}$, is given by

$$C_{D0}^{3/4} = 0.21 / S \Rightarrow C_{D0} = (0.21 / S)^{3/4}. \tag{29}$$

Equation (27) can be rewritten as

$$f\,x = V(1 - w/V) = V\,(1 - 1/6) = V(5/6). \tag{30}$$

The Strouhal number $S = (V/x)\ (5/6)\ (h/V) = (5/6)(h/x) = (5/6)\ (1/4.5)$ , and therefore, $C_{D0}$ is

$$C_{D0}^{3/4} = 0.21 / S = 1.134 \tag{31}$$

Thus for Reynolds number, $R_d^* \in \left(10^3, 10^6\right)$, the above equation gives

$C_{D0} = 1.182.$

Transition from laminar to turbulent flow causes an appreciable change in $C_{D0}$ value. Since this depends on factors like turbulence in wind, surface type, mechanical vibrations, it will be assumed that the transition occurs when $Re_{cr} = 3*10^5$ is reached. The value of $C_{D0}$ then falls to 0.3. The value of $C_{D0}$ increases for higher Reynolds number but was assumed to be constant and equal to 0.3. For Reynolds number $R_d^* \in (1,1000)$, $C_{D0}$ was linearly interpolated and approximated by :

$$C_{D0} = -\ 0.113581 R_d + 12.5401 \qquad (31')$$

Mach Number Effect

For bluff bodies, like a cylinder lying across the flow, as the Mach number increases, there is an appreciable change in stagnation pressure, while the base pressure remains unchanged. The drag due to the nose pressure is adjusted by a factor $(1 + 0.25 M^2)$ whereas the drag due to the base pressure remains unchanged. For a plate it has been shown that 50% of the drag is from base and 50% from nose. For the cylinder we do not have such numbers and the estimate used are based on the flat plate. Thus

$$C_{D0} = C_{D0}\left[0.5 + 0.5\left(1 + 0.25M^2\right)\right] = C_{D0}\left[1.0 + 0.125M^2\right] \qquad (32)$$

which is valid up to the transonic region.

Cross Flow Principle

The cross flow principle, states that the fluid dynamic pressure forces on a bluff body corresponds to the velocity component normal to cylindrical axes, which is $Vsin(\alpha)$, as shown in Figure 8.

The following is then valid up to the critical Reynolds number $3*10^5$, if the cross-sectional area $S = d\,b$ is the frontal area for which $C_{Dbasic}$ is defined. Then

$$C_N = N / QS = C_{Dbasic}\ sin2\alpha$$

and

$$\begin{cases} C_D = C_{Dbasic} \sin^3 \alpha \\ C_L = C_{Dbasic} \sin^2 \alpha \cos \alpha \end{cases}$$

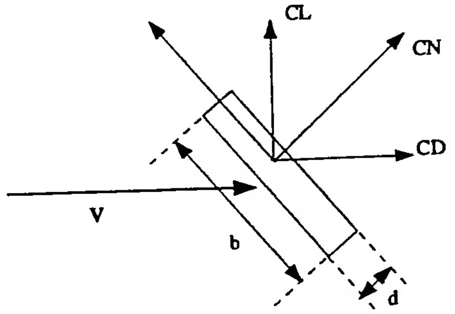


Figure 8. Effective Velocity

These equations are valid for sub-critical Reynolds number $\mathrm{Re}_{cr} = 3*10^5$. Adding $\Delta\ C_D = \pi C_f$ to $C_{Dbasic}$ improves the above prediction. The cross-flow principle is not valid for super-critical conditions.

Reference Area Correction

The reference area for a missile is defined in terms of its cross-sectional area $S'=\pi d^2/4$. Therefore,

$$\begin{cases} C_N = C_N(S/S') = C_{Dbasic}(S/S')\sin^2\alpha \\ C_D = C_{Dbasic}(S/S')\sin^3\alpha \\ C_L = C_{Dbasic}(S/S')\sin^2\alpha\cos\alpha \end{cases} \qquad (33)$$

are the correct aerodynamic coefficients for the missile.

In summary, $C_{D0}$ calculation involves :

(a) Finding $C_{D0}$ based on Reynolds number

(b) Adjusting for compressibility

(c) Using cross-flow correction to get final $C_{D0}$

(d) Correcting the reference area.

Table 4 gives the aerodynamic coefficients for various values of alpha trim and various Mach numbers as obtained from DATCOM and the above prediction techniques. For a reference speed of Mach = 0.8, the plots of lift and drag coefficients over a +- 90 degree range of angle of attack are shown in Figure 9 (for other velocities see the appendix). The rather unconventional maneuver is illustrated in Figure 10, showing different possible values for attitude and air flow direction. Since the angle of attack can reach values greater than 90 degrees, its equivalence to the computed interval is given in Figure 11.

Table 4. Aerodynamic Stability Derivatives

| | **ALPHA TRIM = 10 deg. (DATCOM)** | | | |
|---|---|---|---|---|
| Coeff. | M#=0.3 | M#=0.6 | M#=0.8 | M#=2.0 |
| CNα | 16.094 | 13.212 | 10.875 | 10.015 |
| CNδe | 12.307 | 12.485 | 12.124 | 6.520 |
| CMα | -48.106 | -46.822 - | 44.777 | -10.279 |
| CMδe | -123.266 | -124.985 | -121.278 | -67.076 |
| | **ALPHA TRIM = 40 deg. (DATCOM)** | | | |
| CNα | 6.578 | 44.112 | 52.443 | 31.581 |
| CNδe | 5.690 | 6.452 | 6.967 | 4.864 |
| CMα | -68.927 | -103.362 | -104.679 | -58.786 |
| CMδe | -56.992 | -64.572 | -69.683 | -50.008 |
| | **ALPHA TRIM = 80 deg. (HIGH ALPHA)** | | | |
| CNα | 10.796 | 11.156 | 11.529 | 0.0 |
| CNδe | 0.0 | 0.0 | 0.0 | 0.0 |
| CMα | 0.0 | 0.0 | 0.0 | 0.0 |
| CMδe | 0.0 | 0.0 | 0.0 | 0.0 |

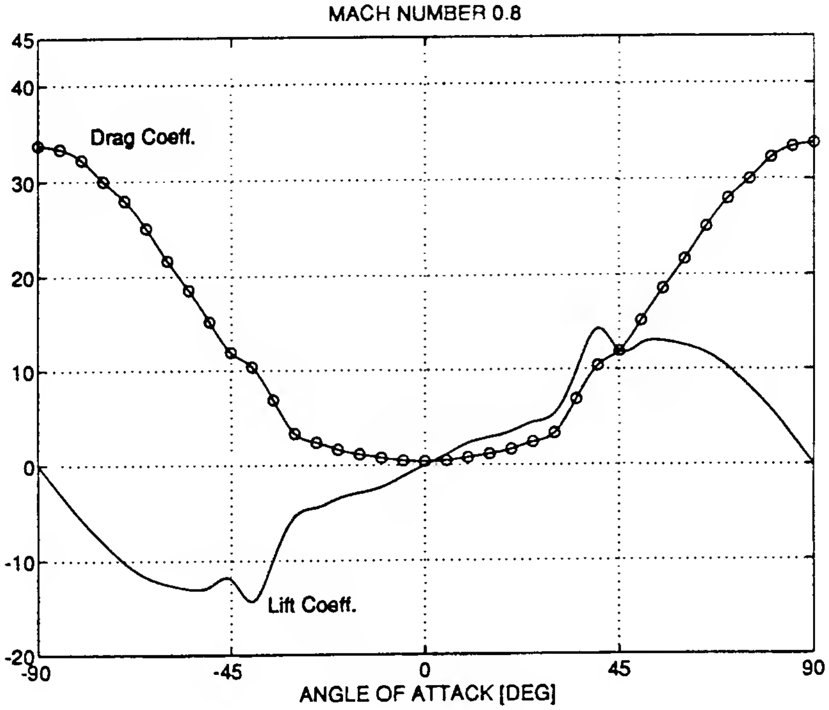


Figure 9. Lift and Drag Coefficients Curves

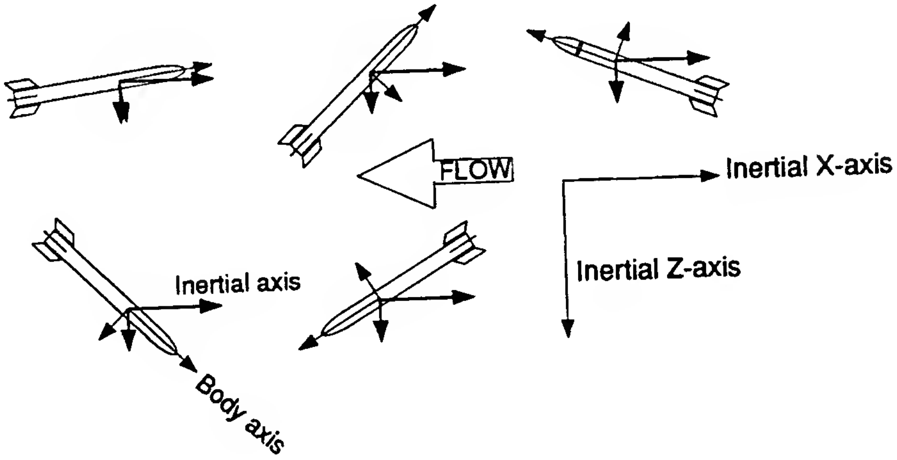


Figure 10. Angular Relationships during Maneuver

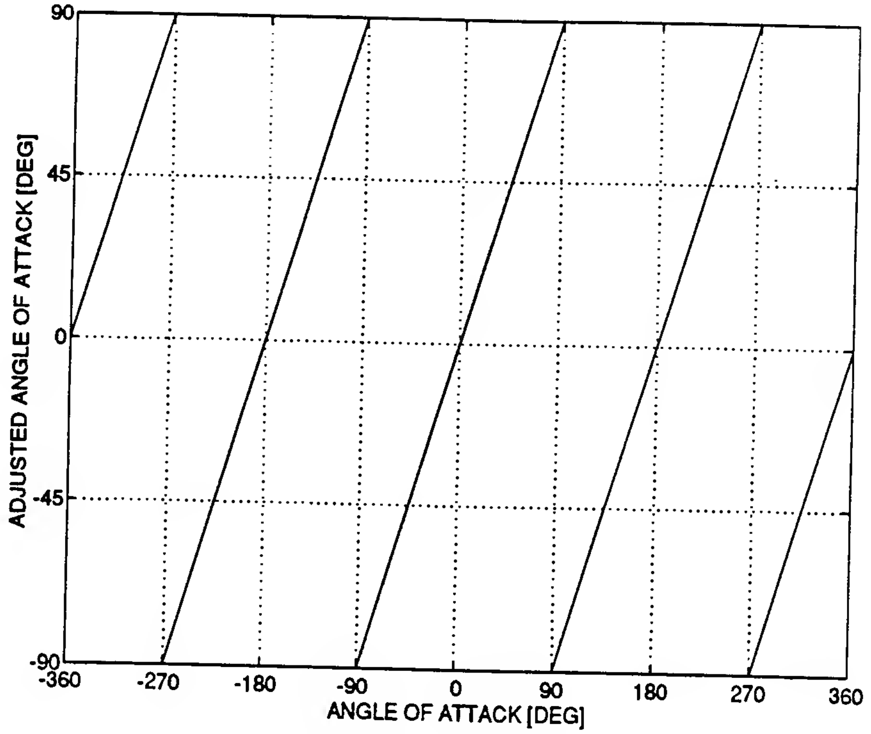


Figure 11. Adjusted versus Actual Angle of Attack

## 3.3. Low Angle of Attack Model

In the introduction, the test maneuver and the autopilot design were confined to the longitudinal plane. Due to the wide range of dynamic pressure experienced by the vehicle, the modelling was divided into two parts. A simplified rigid body linearized motion at angles of attack below stall, and a combined pitch rotation and point mass translation above stall.

In the low angle of attack region, the speed was assumed constant and a typical short period approximation was used [1], [26]. In equation form we have

$$\begin{cases} \dot{\alpha} = \dfrac{-QS}{mV_T} C_{N\alpha}\alpha + q - \dfrac{QS}{mV_T} C_{N\delta}\delta - \dfrac{T_{RCS}}{mV_T} u_T \\ \dot{q} = \dfrac{QSL_{REF}}{I_y} C_{m\alpha}\alpha + \dfrac{QSL_{REF}}{I_y} C_{m\delta}\delta + \dfrac{T_{RCS}L_{RCS}}{I_y} u_T \end{cases} \tag{34}$$

which becomes, in dimensional form

$$\begin{cases} \dot{\alpha} = Z_\alpha \alpha + q + Z_\delta \delta + Z_T u_T \\ \dot{q} = M_\alpha \alpha + M_\delta \delta + M_T u_T \end{cases} . \tag{34'}$$

$T_{RCS}$ is the thrust provided by reaction jet and $L_{RCS}$ is the jets moment arm. The actuators are modelled by linear first order systems with time constants $\tau_d$ and $\tau_u$ for the elevator and RCS respectively. In equation form

$$\begin{cases} \tau_\delta \dot{\delta} = \delta + \delta_c \\ \tau_u \dot{u}_T = u_T + u_{Tc} \end{cases} . \tag{35}$$

Since standard measurements and commands in missile autopilots include accelerometers and rate gyros, Eqs. (34') and (35) can be rewritten in state space form by introducing the normal load factor $N_z$ as a state variable replacing the angle of attack. From kinematics we have

$$N_z = \frac{a_z}{g} = \frac{V_T}{g}(\dot{\alpha} - q) \tag{36}$$

By defining $x^T = [N_z \quad q \quad \delta \quad u_T]$ and $u^T = [\delta_c \quad u_{Tc}]$ we obtain

$$\begin{bmatrix} \dot{N}_Z \\ \dot{q} \\ \dot{\delta} \\ \dot{u}_T \end{bmatrix} = \begin{bmatrix} Z_\alpha & \frac{V_T Z_\alpha}{g} & -\frac{V_T Z_\delta}{g\tau_\delta} & -\frac{V_T Z_T}{g\tau_u} \\ \frac{gM_\alpha}{V_T Z_\alpha} & 0 & M_\delta - \frac{M_\alpha Z_\delta}{Z_\alpha} & M_T - \frac{M_\alpha Z_T}{Z_\alpha} \\ 0 & 0 & -\frac{1}{\tau_\delta} & 0 \\ 0 & 0 & 0 & -\frac{1}{\tau_u} \end{bmatrix} \begin{bmatrix} N_Z \\ q \\ \delta \\ u_T \end{bmatrix} + \begin{bmatrix} \frac{V_T Z_\delta}{g\tau_\delta} & \frac{V_T Z_T}{g\tau_u} \\ 0 & 0 \\ \frac{1}{\tau_\delta} & 0 \\ 0 & \frac{1}{\tau_u} \end{bmatrix} \begin{bmatrix} \delta_c \\ u_{Tc} \end{bmatrix}$$

or, in standard form

$$\dot{x} = Ax + Bu \tag{37}$$

with the output vector being equal to the state vector.

The time rate of change of the flight path angle is computed from the normal acceleration, provided the angle of attack is small, as

$$\dot{\gamma} = -\frac{a_z}{V_T} \tag{38}$$

and the inertial velocity components are

$$\begin{cases} \dot{X} = V_T \cos\gamma \\ \dot{Z} = V_T \sin\gamma \end{cases} \tag{39}$$

## 3.4. High Angle of Attack Model

The model for the post stall region must account for the dynamic pressure variation, which is characterized by large excursions in angle of attack, attitude, flight path angle, and velocity. In this region, aerodynamic control becomes negligible and the only effective actuators are the reaction jets. Such control must rotate the vehicle counteracting the pitching moment produced by the normal force, which tends to oppose the motion. Considering a pure rotational attitude motion, we have

$$I_y\dot{q} = QSC_N L_{cp} + L_{RCS}T_{RCS}u_T \tag{40}$$

where $L_{cp}$ is the distance between center of pressure (as computed from the assumed cylindrical configuration of section 3.2.) and the center of mass. The value of $L_{cp}$ will be a design parameter for future studies, when a more detailed aerodynamic model will be available.

The velocity variation is taken into account by adding a two-dimensional point mass model, which in the inertial reference is given by

$$\begin{cases} m\ddot{X} = -L\sin\gamma - D\cos\gamma + T_E\cos\theta - T_{RCS}\sin\theta \\ m\ddot{Z} = -L\cos\gamma + D\sin\gamma - T_E\sin\theta - T_{RCS}\cos\theta + mg. \\ \gamma = \theta - \alpha \end{cases} \tag{41}$$

The flight path angle is also related to the inertial velocity components from

$$\gamma = -\tan^{-1}\left(\frac{\dot{Z}}{\dot{X}}\right) \tag{42}$$

differentiating Eq. (42) with respect to time yields

$$\dot{\gamma} = \frac{\dot{Z}\ddot{X} - \ddot{Z}\dot{X}}{\dot{X}^2 + \dot{Z}^2}$$

therefore the normal acceleration is

$$a_z = \ddot{X}\sin\theta + \ddot{Z}\cos\theta \tag{43}$$

In Eq. (41), one of the force components is the main engine thrust $T_E$. In this phase, the main engine is fired to recover the speed loss experienced by the vehicle during the rotation. The main engine thrust is set to be a function of the attitude angle in an open loop fashion, and it is operational within preset values of the overall speed.

## 3.5. Acquisition of Steady State

At the end of the maneuver, the vehicle enters the low angle of attack regime with full dynamic pressure recovery. The reaching of the desired steady state requires specific values for the attitude, flight path angle, and consequently the angle of attack itself. The point mass equations are still valid and the rotational motion equation must now include the flight path angle. This model for the final phase of the trajectory is used, instead of simple short period, to account for flight path and velocity variations still present in this phase. Keeping Eq. (41), we now use

$$\dot{q} = M_\alpha \alpha + M_\delta \delta + M_T u_T \tag{44}$$

in place of Eq. (40). Replacing $\alpha$ with $\theta - \gamma$ in Eq. (44) yields

$$\dot{q} = M_\alpha \theta - M_\alpha \gamma + M_\delta \delta + M_T u_T \tag{45}$$

where the flight path angle can now considered as an additional input or state depending on the approach used for the autopilot design.

Since during this phase of the maneuver the vehicle is flying at an attitude and flight path angles close to 180 degrees, the small perturbation Eq. (45) leads to a mathematical singularity in that the controller does not distinguish between 180 and zero degrees. To eliminate the problem, a simple 180-degree roll maneuver would be sufficient, however because our work was limited to the longitudinal plane, an alternate approach was taken consisting of the introduction of the desired steady state (trim) values in Eq. (45).

Defining the desired set of trim point values as $[\theta_0 \quad \gamma_0 \quad \delta_0 \quad u_{T0}]$, the perturbation variables can be written as $\theta = \Theta - \theta_0$, $\gamma = \Gamma - \gamma_0$, $\delta = \Delta - \delta_0$, and $u_T = U_T - u_{T0}$, (note that zero steady state pitch rate is assumed). Therefore Eq. (45) becomes

$$\dot{X} = \begin{bmatrix} \dot{\Theta} \\ \dot{q} \\ \dot{\Gamma} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ M_\alpha & 0 & -M_\alpha \\ -Z_\alpha & 0 & Z_\alpha \end{bmatrix} \begin{bmatrix} \Theta - \theta_0 \\ q \\ \Gamma - \gamma_0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ M_\delta & M_T \\ -Z_\delta & Z_T \end{bmatrix} \begin{bmatrix} \Delta - \delta_0 \\ U_T - u_{T0} \end{bmatrix} \qquad (46)$$

where $X = [\Theta, q, \Gamma]^T$ and $\dot{X} = \dot{x}$ .

If the flight path angle is considered as input rather than state, we have

$$\begin{bmatrix} \dot{\Theta} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ M_\alpha & 0 \end{bmatrix} \begin{bmatrix} \Theta - \theta_0 \\ q \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ M_\delta & M_T \end{bmatrix} \begin{bmatrix} \Delta - \delta_0 \\ U_T - u_{T0} \end{bmatrix} + \begin{bmatrix} 0 \\ -M_\alpha \end{bmatrix} (\Gamma - \gamma_0). \qquad (46')$$

In summary, Eqs. (41) and (46) or Eqs. (41) and (46') represent the missile model in the final phase of steady state acquisition.

### 4. Autopilot Design

This section describes the design of the pitch channel autopilot using variable structure control. The control logic is broken down according to the phase of flight, resulting in a structure that includes the low alpha component, the post stall rotation, and the final steady state acquisition.

## 4.1. Objectives

The main objective of the present work was the synthesis of a pitch autopilot that uses a combination of aerodynamic and reaction jet control to achieve a 180-degree longitudinal heading maneuver in the vertical plane. As described in the previous sections, the chosen maneuver consists of a full turn reversal, which requires for being successful, high angles of attack, loss of dynamic pressure, and the acquisition of a final state characterized by a straight level flight with speed equal to the initial velocity. Our concentration has been on the pitch autopilot, assuming no interaction among the three channels. Additional effort is suggested and has been proposed [25], to study the full spatial implications of post stall maneuvering.

The flight condition chosen for the point design corresponds to Mach 0.8 and altitude of 10,000 feet. The design goal was a controller structure, insensitive to the uncertainty of the values of the aerodynamic coefficients and different values of reaction jets thrust, because this will limit gain scheduling.

The autopilot design was found by using the VSC technique described in section 2. The control structure was changed as function of the angle of attack. The result was a three-phase design. The first phase, (Phase I) begins after the initial command and ends when stall is reached. The second phase, (Phase II) is flown at angles of attack above stall. This phase encompasses the main part of the trajectory and involves a model following of the attitude, as well as the open loop boost section for the rotation of the velocity vector and its magnitude recovery. The final phase, (Phase III) is again below stall and it is implemented as tracking of attitude and flight path set points.

In addition to having a different control logic, the three phases differ in their implementation. While the first and third use both reaction jet and elevator; the second phase control relies on the use of reaction jet and main engine thrust.

## 4.2. Phase I Autopilot

The vehicle's aerodynamic coefficients, as obtained from DATCOM, for a trim angle of attack of 10 degrees, at various speed are shown in Table 5.

Table 5. Aerodynamic Coefficients from DATCOM

| **Mach** | **0.3** | **0.6** | **0.8** | **1.0** | **1.2** | **1.4** | **1.8** | **2.0** |
|---|---|---|---|---|---|---|---|---|
| $C_{N\alpha}$ | 16.1 | 13.2 | 10.9 | 10.2 | 10.0 | 7.83 | 15.7 | 16.6 |
| $C_{N\delta}$ | 12.3 | 12.5 | 12.1 | 12.6 | 11.6 | 9.43 | 7.62 | 6.52 |
| $C_{m\alpha}$ | -48.1 | -46.8 | -44.8 | -48.8 | -41.8 | -22.0 | -15.1 | -10.3 |
| $C_{m\delta}$ | -123.3 | -125.0 | -121.3 | -125.6 | -118.2 | -96.26 | -78.10 | -67.10 |

Using the values corresponding to Mach 0.8, the system's matrices in Eq. (37) become

$$A = \begin{bmatrix} -.1647 & -4.4082 & 884.6 & 1109.1 \\ 7.5829 & 0 & -53.27 & 47.87 \\ 0 & 0 & -180 & 0 \\ 0 & 0 & 0 & -500 \end{bmatrix} \quad B = \begin{bmatrix} -884.6 & -1109.1 \\ 0 & 0 \\ 180 & 0 \\ 0 & 500 \end{bmatrix}$$

indicating a very lightly damped short period with natural frequency $\omega = 5.78$ rad/sec and damping factor $\zeta = 0.01424$. The autopilot objective in this phase is the tracking of a g-command to achieve a rapid pitching motion. In this respect, the autopilot does not differ from other traditional missile autopilots [3], [26]. The main reason for choosing a g-command structure strategy was to preserve the control logic available from standard designs and used in parts of the flight envelope not requiring post stall maneuvering, nor reaction jet.

In order to achieve perfect tracking of the load factor, an integrator was added to the state vector, which then becomes $x_{aug}^T = [\int N_z dt \quad N_z \quad q \quad \delta \quad u_T]$ leading to a VSC-PI control structure. The augmented system dynamics are

$$\dot{x}_{aug} = A_{aug} x_{aug} + B_{aug} u. \tag{47}$$

From (15), the controller structure is

$$u = -L x_{aug} - \rho \frac{N x_{aug}}{\| M x_{aug} \| + \delta}. \tag{48}$$

The values of the gain matrices can be found in the Appendix A1 and were obtained using the following values for the design parameters, with $\delta = 0.01$

$$\begin{cases} Q = diag[45 \quad .1 \quad .1 \quad .1 \quad 40] \\ \Lambda_3^* = diag[-10 \quad -10] \\ \rho = \begin{bmatrix} \rho_\delta = 1 & 0 \\ 0 & \rho_{uT} = 10 \end{bmatrix} \end{cases} . \tag{49}$$

The selection of weightings in Eq. (49) was based on obtaining a fast load factor response and at the same time on keeping the control efforts within their limitations. The matrix G which defines the sliding surface $s = G x_{aug}$ is also found in the Appendix A1. Figure 12 shows the block diagram implementation of the autopilot.

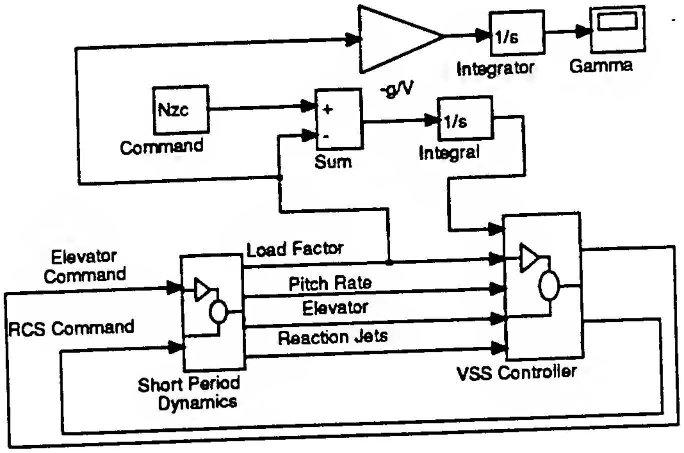


Figure 12. Phase I Autopilot

The performance of the autopilot to a step command in load factor are shown in Figure 13, Ref. [1] has a comparison with a standard autopilot. The control effort, angle of attack and attitude responses are plotted in Figure 14. Considering a -5g command as the representative one, the vehicle reaches stall in less than 0.4 seconds. At this point, the autopilot's logic switches to Phase II.

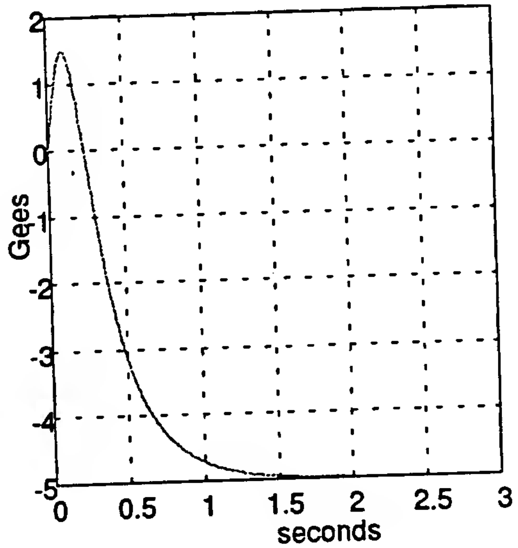


Figure 13. G-Command Response

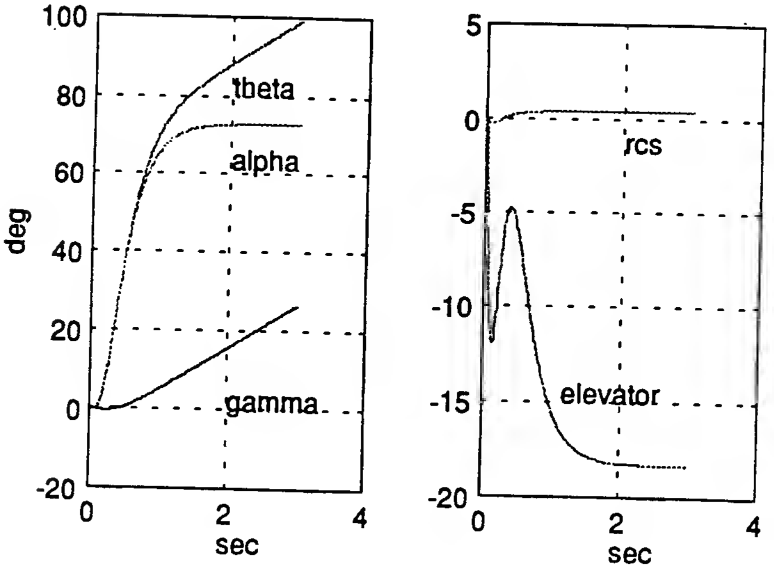


Figure 14. Performance during Phase I

In the simulation section, the reaction jet actuation has a on-off structure with values -1, 0, and 1 corresponding to the system being on the sliding surface ($u_T = 0$) or off of it. When $s = Gx_{aug} < 0$, then $u_T = -1$, otherwise $u_T = +1$. The stability of the system is ensured by the fact that the reaction jet equivalent control $u_{Teq}$ remains bounded between +1 and -1 [27].

### 4.3. Phase II Autopilot

This phase is characterized by a pure rotation in pitch generated by RCS alone. In this phase, the vehicle is operating at angles of attack above stall and the aerodynamic forces and moments are not considered useful in controlling the missile. The autopilot is designed to operate as a model following VSS controller in feedback, with an outer loop consisting of the main engine firing according to the values of attitude and speed.

The main objective of the autopilot in this phase is the rotation of the vehicle and compensation of dynamic pressure, which drops considerably due to speed reduction associated with drag increase at high angles of attack. To recover dynamic pressure, the main engine is fired in an open loop fashion at a predetermined attitude chosen, nominally, θ = 120 degrees. The engine remains in a boost phase mode until the final desired speed (here assumed equal to the initial one) is achieved as well as the appropriate direction of flight. The equations of motion used for the controller synthesis are Eqs. (40) and (41). VSC is applied to Eq. (40), that can be written in state space form as

$$\dot{x} = \begin{bmatrix} \dot{\theta} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \dfrac{L_{RCS} T_{RCS}}{I_y} \end{bmatrix} u_T + \begin{bmatrix} 0 \\ \dfrac{Nd}{I_y} \end{bmatrix} = Ax + Bu_T + D \qquad (50)$$

Here $N$ is the normal force acting on the vehicle and $d$ is the moment arm (the contribution of the axial force was considered negligible at post stall angles of attack). Globally, the torque due to normal force produces a resisting effect that can be treated as a disturbance $D$. To perform a point design of the autopilot, a worst case approach was used and $D$ was bounded by the value corresponding to a cylinder set perpendicular to the flow, where the drag assumes its maximum value. With this, the B and D terms in Eq. (50) become

$$B = \begin{bmatrix} 0 \\ 31.049 \end{bmatrix}, D = \begin{bmatrix} 0 \\ -10 \end{bmatrix}$$

Next, VSC was applied using a model following approach [9], [12], by specifying a desired model for the pitch rotational dynamics. The model was chosen based on the +1, -1 limitations of the reaction jet, and it is given by

$$\dot{x}_m = A_m x_m + B_m u_m = \begin{bmatrix} 0 & 1 \\ -178 & -24 \end{bmatrix} x_m + \begin{bmatrix} 0 \\ 178 \end{bmatrix} u_m \tag{51}$$

from (50) and (51), the error dynamics $e = x_m - x$ are given by

$$\dot{e} = A_m e + [A_m - A]x + B_m u_m - D - Bu_T \tag{52}$$

a general form for the control law, for perfect model following, can be written as

$$u_T = u^{NL} + K_1 x + K_2 u_m + K_3 = u^{NL} + u^L \tag{53}$$

with

$$\begin{cases} K_1 = B^*[A_m - A] \\ K_2 = B^* B_m \\ K_3 = B^* D \end{cases} \tag{54}$$

and $B^* = (B^T B)^{-1} B^T$ is the appropriate pseudoinverse of $B$.

In the present case, the implementation of the control law Eq. (53) is limited to the nonlinear component, which is given by

$$\begin{cases} u_T = u^{NL} = sgn(s) \\ s = Ge = [7 \quad 1]e. \end{cases} \tag{55}$$

The linear component of the control law is nothing but $u_{eq}$, the controller structure during sliding. The choice of G was made to ensure a desirable response during sliding.

Phase II starts with initial conditions corresponding to the state variables final conditions from Phase I. This phase lasts about 0.5 seconds and shows excellent performance. The results in terms of model following and control effort are presented in Figures 15 and 16, and the structure of the autopilot is shown in Figure 17.

The attitude angle goes from 40 degrees to 180 degrees. At 120 degrees, the main engine is turned on and its effect is evident from the speed and flight path angle time histories included in the simulation section. The engine is shut down when the speed equivalent to Mach 0.8 is recovered, although other choices can be made to decrease/increase the boost phase. At the end of Phase II, the angle of attack is below stall and Phase III autopilot becomes active. The above procedure will be shown in the simulation section.

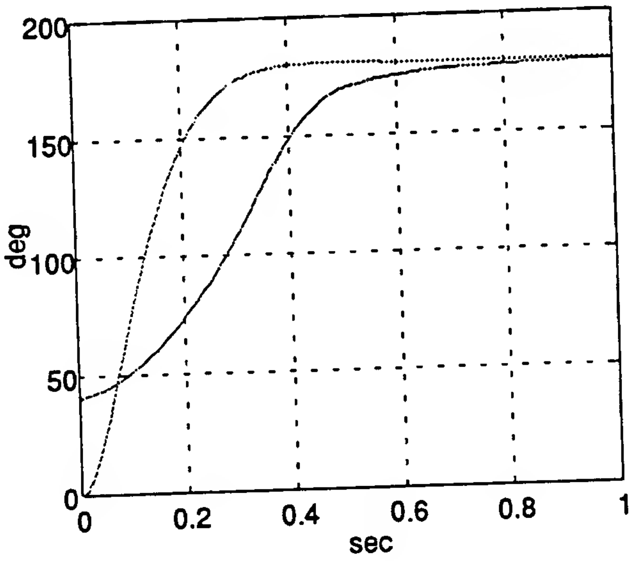


Figure 15. Performance during Phase II

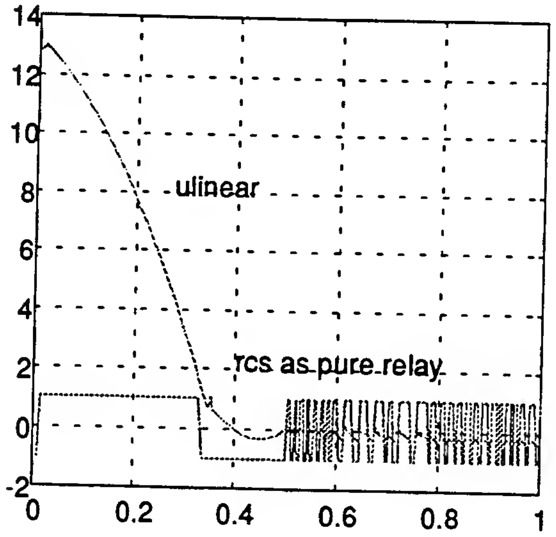


Figure 16. Control Effort during Phase II

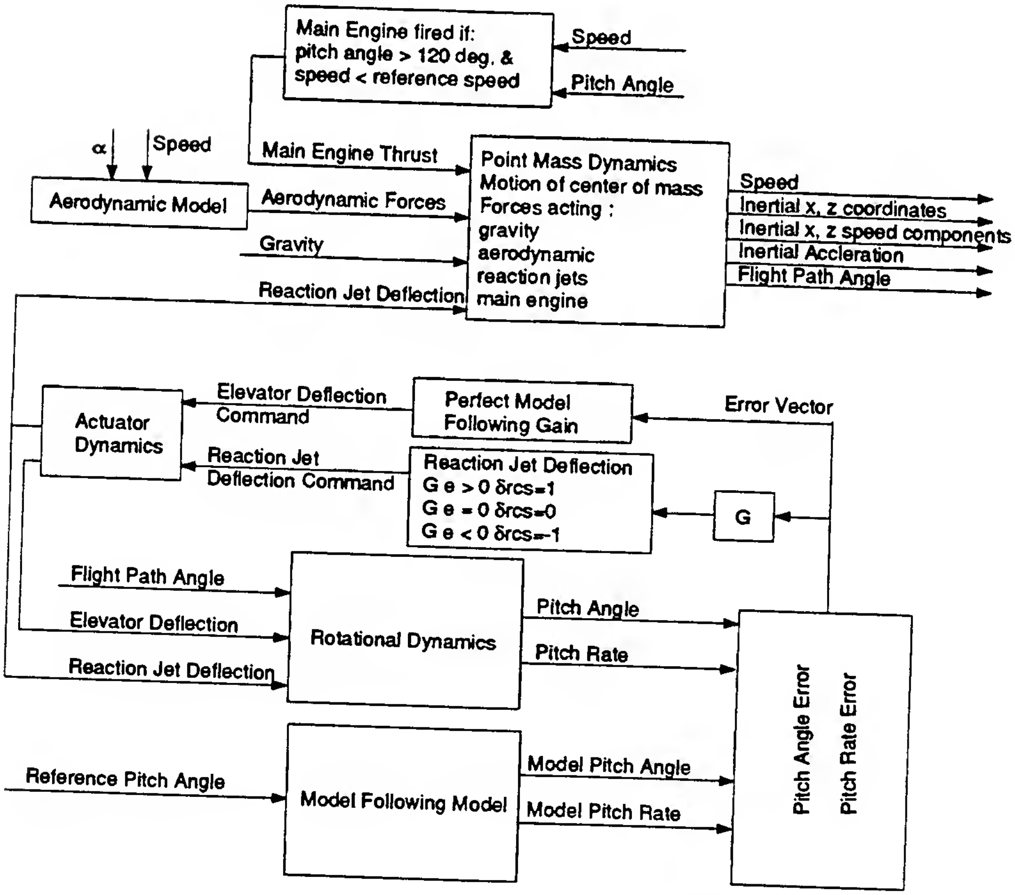


Figure 17. Phase II Autopilot Block Diagram

## 4.4. Phase III Autopilot

This phase involves the task of acquiring and maintaining a set of steady state values for the attitude and flight path of the missile. In our work we established the objective of reaching 180 of flight path and attitude angles, although different values can be set if desired.

The pertinent model for the system is given by Eqs. (41) and (46). Since the flight path angle is required to reach 180 degrees, the rotational equation now contains $\gamma$ as an additional state variable. Using the numerical values from Table 5, and defining the trim conditions $x_0 = [\theta_0, q_0, \gamma_0]^T$ and $u_0 = [\delta_0, u_{T0}]^T$, Eq. (46) now becomes

$$\dot{X} = \dot{x} = A(X - x_0) + B(U - u_0) \tag{56}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -33.4266 & 0 & 33.4266 \\ 0.1647 & 0 & -0.1647 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ -90.5355 & 31.049 \\ 0.1836 & 0.0829 \end{bmatrix}$$

using the VSC model following approach, we define the model dynamics as

$$\dot{X}_m = \dot{x}_m = A_m(X_m - x_{m0}) + B_m(U_m - u_{m0}) \tag{57}$$

where

$$A_m = \begin{bmatrix} 0 & 1 & 0 \\ -16 & -8 & 0 \\ 0 & 0 & -0.5 \end{bmatrix} \quad B_m = \begin{bmatrix} 0 & 0 \\ 16 & 0 \\ 0 & 0.5 \end{bmatrix} \text{ and } U_m - u_{m0} = \begin{bmatrix} \theta_c = 180^o \\ \gamma_c = 180^o \end{bmatrix}$$

The model following controller again operates on the error state, difference between $X$ and $X_m$ and whose dynamics are given by

$$\dot{e} = A_m e + [A_m - A]x + B_m u_m + Ax_0 - Bu \tag{58}$$

the control law has the form

$$u = u^{NL} + K_1(X - x_0) + K_2 x_0 + K_3 U_m = u^{NL} + u^L \tag{59}$$

with

$$\begin{cases} K_1 = B^*[A_m - A] \\ K_2 = B^* A \\ K_3 = B^* B_m \end{cases} \tag{60}$$

if the value of $U_m$ and $x_0$ are the same, as in our case, then the controller becomes

$$u = u^{NL} + K_1(X - x_0) \tag{61}$$

with the gain $K_1$ given in Appendix A1. From Eq. (61) into (58) we obtain the error dynamics as

$$\dot{e} = A_m e - Bu^{NL} \tag{62}$$

the structure of the nonlinear component in Eq. (62) is based on the Eq. (48) and given by

$$u = -Le - \rho \frac{Ne}{\|Me\| + \delta} \tag{63}$$

with the matrix gains above computed similarly to those in Eq. (48) using the following parameters

$$\begin{cases} Q = diag[1 \quad 100 \quad 10] \\ \Lambda_3^* = diag[-3 \quad -3] \\ \rho = \begin{bmatrix} \rho_\delta = .01 & 0 \\ 0 & \rho_{uT} = .01 \end{bmatrix}. \end{cases} \tag{64}$$

In the implementation, the elevator deflection operates following Eq. (63), while the reaction jet operates in an on-off fashion according to the sign of the sliding surface erros $s = Ge$ [27].

A control strategy for the Phase III autopilot based on model following of the attitude only was also derived. In this case, the appropriate equations of motion for the vehicle become (41) and (46'). The flight path angle is then considered as a disturbance to the system, leading to the following rotational equations for the vehicle and the model

$$\begin{bmatrix}\dot{\Theta}\\ \dot{q}\end{bmatrix}=\begin{bmatrix}0 & 1\\ -33.4266 & 0\end{bmatrix}\begin{bmatrix}\Theta-\theta_0\\ q\end{bmatrix}+\begin{bmatrix}0 & 0\\ -90.536 & 31.049\end{bmatrix}\begin{bmatrix}\delta\\ u_T\end{bmatrix}+\begin{bmatrix}0\\ -33.4266\end{bmatrix}\gamma \tag{65}$$

$$\begin{bmatrix}\dot{\theta}_m\\ \dot{q}_m\end{bmatrix}=\begin{bmatrix}0 & 1\\ -36.4 & -12\end{bmatrix}\begin{bmatrix}\theta_m\\ q_m\end{bmatrix}+\begin{bmatrix}0\\ 36.4\end{bmatrix}u_m.$$

The procedure is similar to the previous one leading to a controller of the form

$$\begin{cases}\delta = K_x(X-x_0)+K\gamma(\Gamma-\gamma_0)+K_{um}(U_m-u_{m0})\\ u_T=\begin{cases}1 & (s>0)\\ 0 & (s=0)\\ -1 & (s<0)\end{cases}\end{cases} \tag{66}$$

with $s = Ge = 7(\theta_m - \theta) + (q_m - q)$, where the gains in Eq. (66) are reported in Appendix A1. Note that, with this second approach, the elevator is responsible for providing model following, while the reaction jet control gives robustness to parameter variations and disturbances. The block diagrams of Phase III autopilots are shown in Figures 18 and 19.

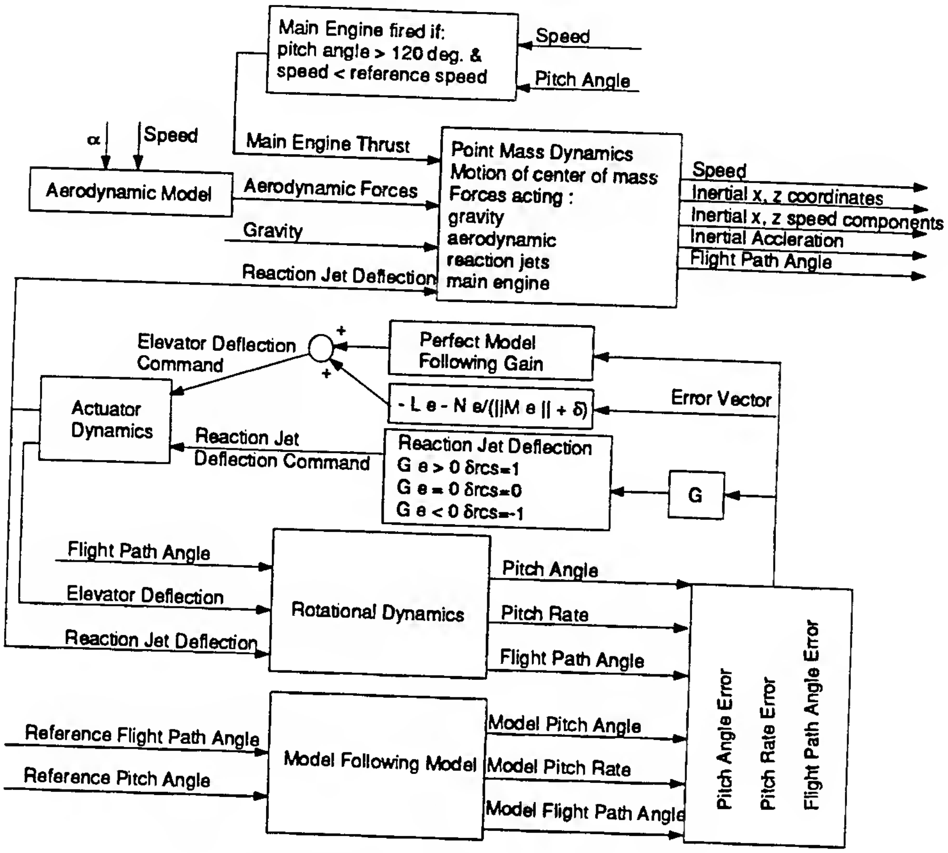


Figure 18. Phase III Autopilot Block Diagram (method 1)

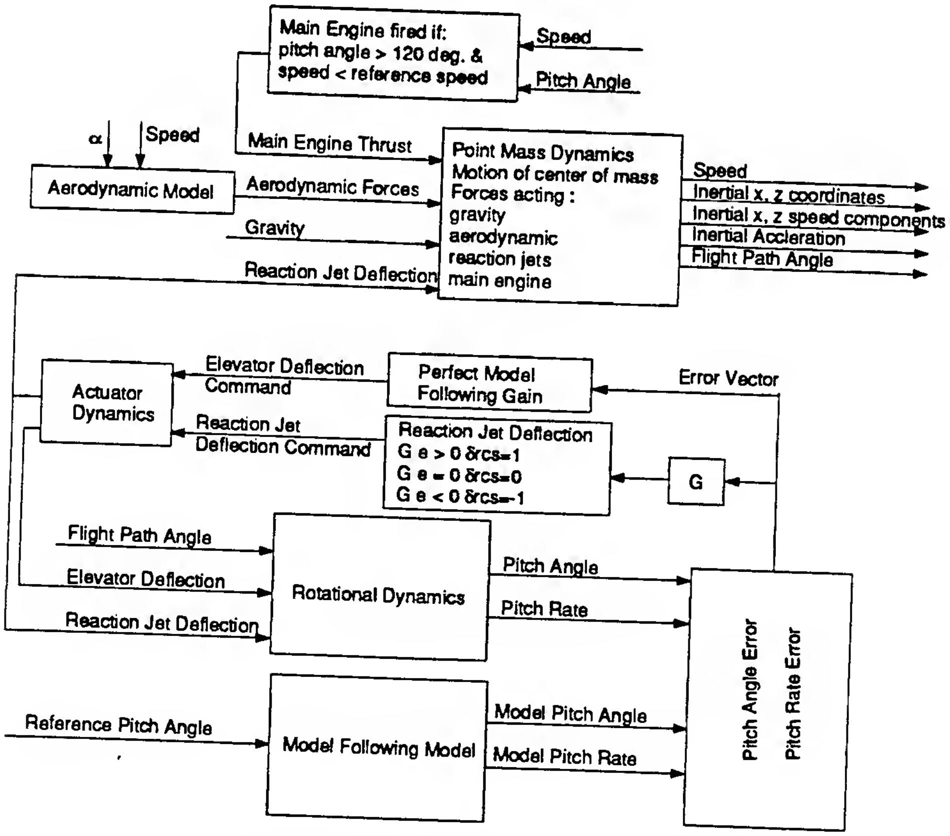


Figure 19. Phase III Autopilot Block Diagram (method 2)

## 5. Simulation

This section presents the simulation results relative to the entire maneuver for the nominal case as well as sensitivity analysis in terms of some of the vehicle's parameters such as reaction jet's thrust, initial speed, and attitude angle at main engine firing time.

### 5.1. Description

The autopilot logic described in section 4, and whose diagram is summarized in Figure 20, was tested with a nonlinear simulation code developed in the Matlab®-Simulink™ environment, and enclosed in this report as deliverables. The main objectives of the simulation were:

1. To validate the autopilot performance in the chosen scenario.
2. To test the robustness of the methodology by using fixed controller gains instead of gains scheduled with angle of attack and Mach number variation.
3. To test the robustness of the methodology with respect to parameter variations such as reference speed, initial time of main engine firing, and reaction jet reference thrust.

It is important to note that, during the simulation, the aerodynamic forces and moments are not constant, but vary according to Figure 9, as functions of the angle of attack.

One of the problems encountered in the development of the simulation was the choice of integration routine, since the results were found to be quite sensitive to the integration technique and step size selection. The most satisfactory results were obtained using a third order Runge-Kutta method with minimum step size $\Delta T = 0.001$.

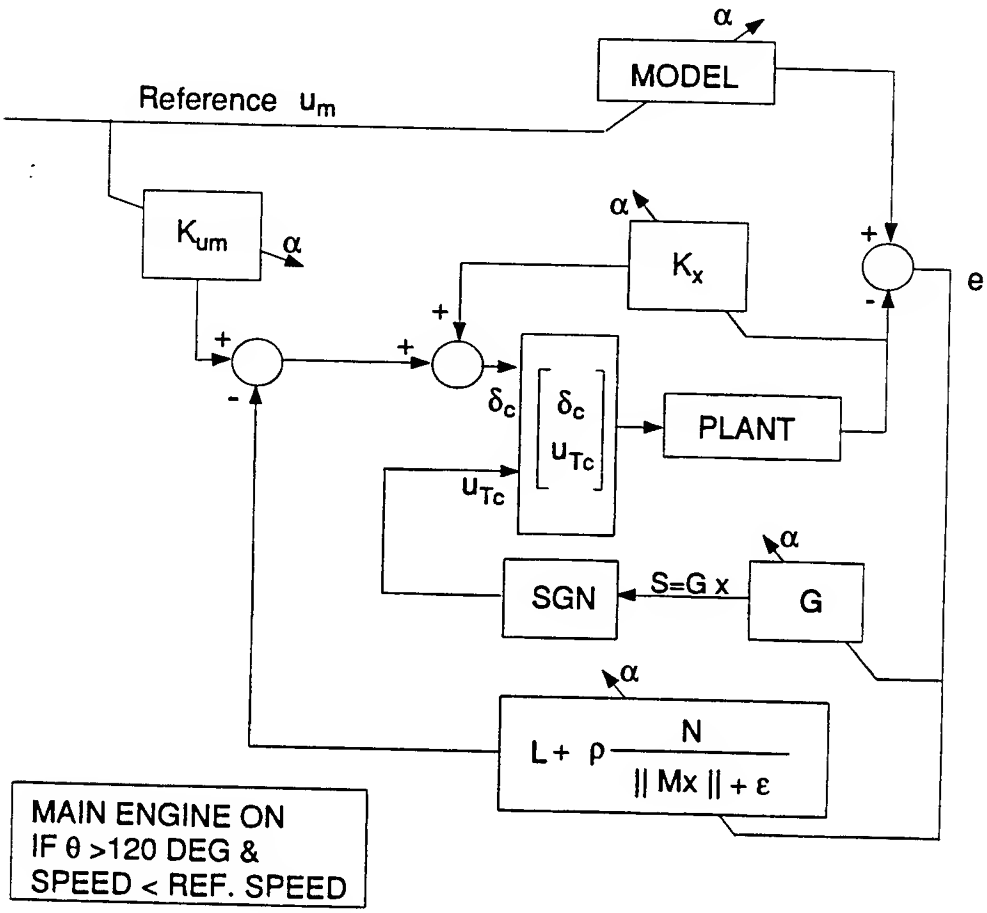


Figure 20. Complete Autopilot Schematic

Some of the time histories show slight oscillations due to the choice of step size. Such oscillations can be eliminated by reducing ΔT at the expense of disk space.

## 5.2. Results

### 5.2.1. Nominal Case

Nominal values for the geometric and aerodynamic characteristics of the system are found in Tables 1 and 3. The autopilot gains are given in Appendix A1.

The maneuver begins with a negative 5g $N_{zc}$ command. The missile pitches up reaching stall withing 0.278 seconds from the beginning of the maneuver. The motion is essentially in pitch with no appreciable change in flight path angle as seen from Figure 21 where pitch, flight path, and angle of attack are plotted. The load factor at stall has basically reached the commanded value of five times the gravity acceleration as shown in Figure 22, where the top curve corresponds to an integration using a step size of 0.001 seconds and the bottom curve to a step size of 0.00005 seconds. The control activity are shown in Figures 23 and 24.

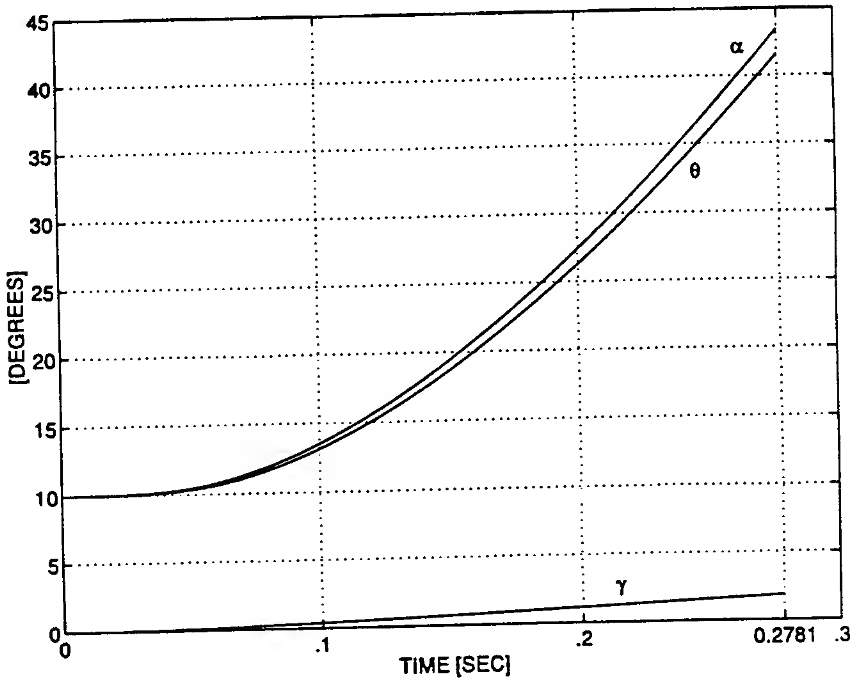


Figure 21. Angular Behavior below Stall

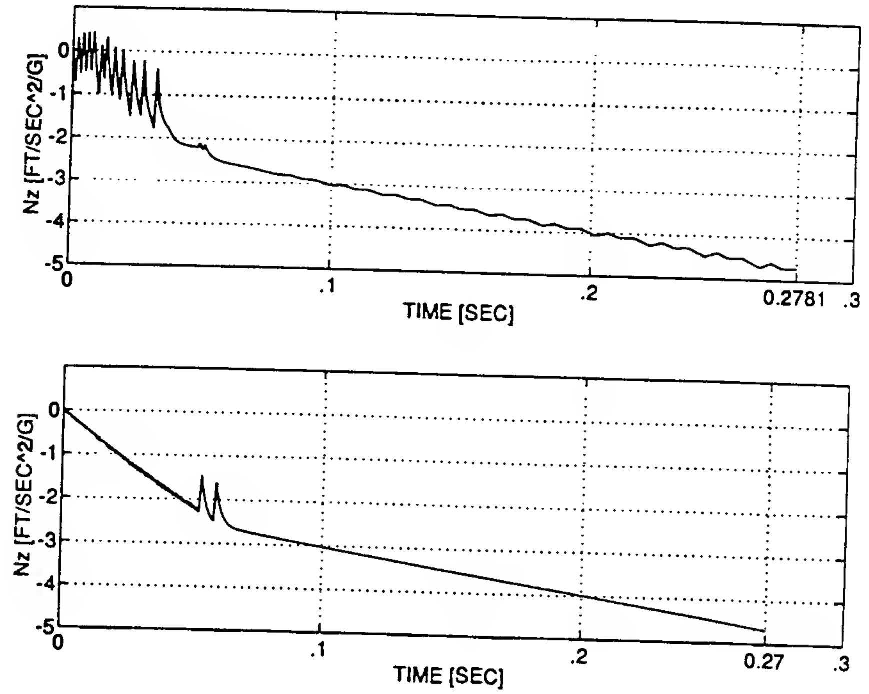


Figure 22. Load Factor Response below Stall

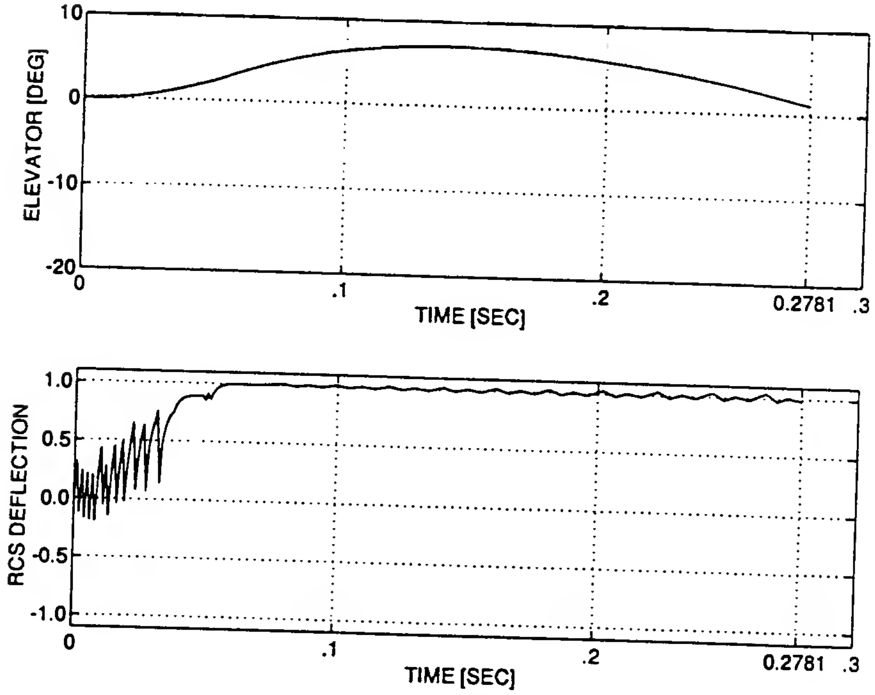


Figure 23. Control Activity during Phase I

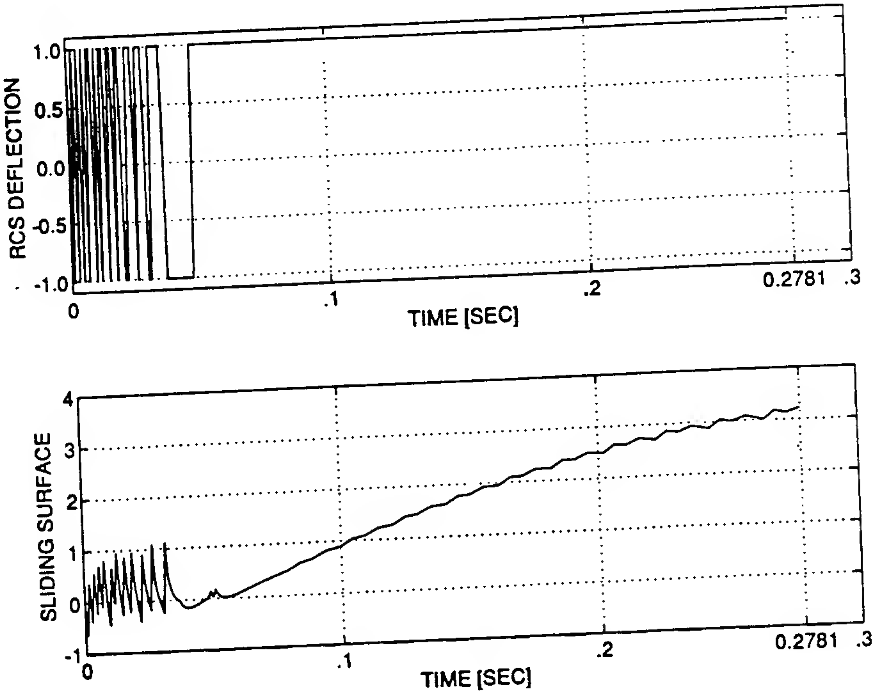


Figure 24. Reaction Jet Activity (On-Off) and Sliding Surface during Phase I

The elevator shows a positive deflection which would mean that the missile would be pitching downwards. But the thrust generats a positive pitching moment which offsets the negative pitching tendency. Figure 24 shows the commanded $u_T$ response with no deadband. The sliding surface is never reached in this particular phase because of stall being achieved prior to reaching it.

Once the stalled phase is reached (pre-selecteded to be 40 degrees of angle of attack), the system model changes to the high angle of attack, so does the autopilot structure which becomes a second order model following VSS. Figure 25 shows actual and model attitude and pitch rate denoted by lines 2 and 1 respectively. The missile's responses in attitude, angle of attack and flight path angle are given in Figure 26. In this phase, the main engine is activated when the attitude angle reaches a value of 120 degrees, just prior to 0.5 seconds after the beginning of the maneuver. The effect of the engine is clearly shown in Figure 26.

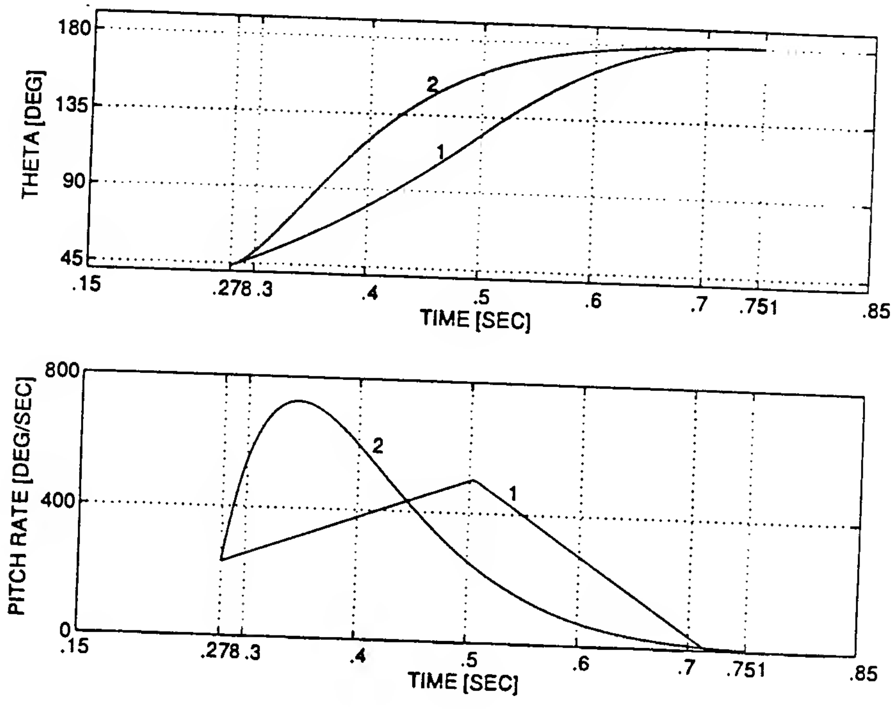


Figure 25. Attitude and Pitch Rate during Phase II

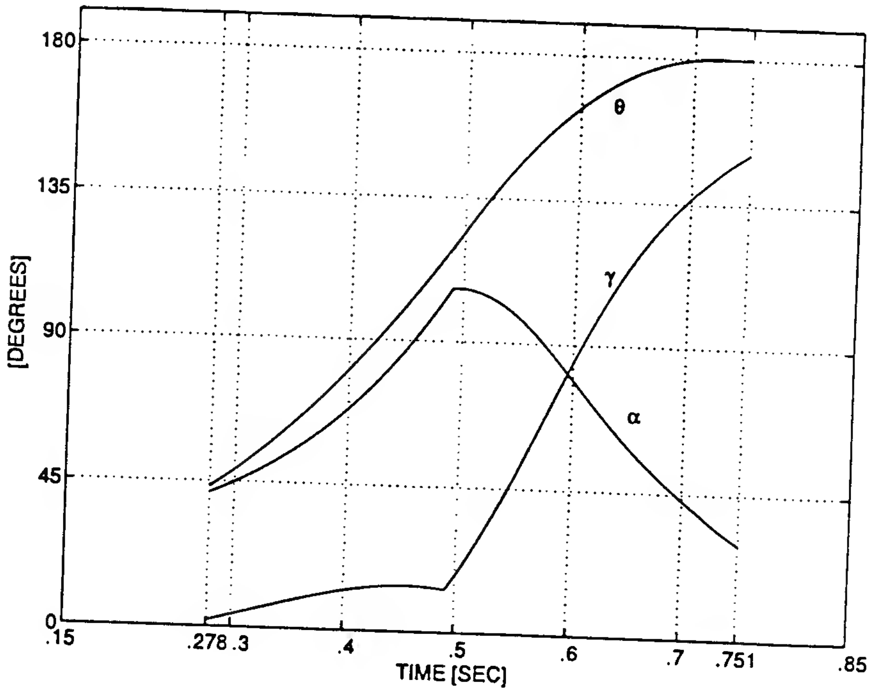


Figure 26. Missile Angular Behavior during Phase II

Figure 27 shows the reaction jet activity during this phase. The reaction jet's thrust is an On/Off system with deadband and has a first order dynamics. Line 2 is the reaction jet working with no deadband, while line 3 shows the behavior with no deadband nor actuator lag. Each one is equally effective in controlling the motion.

The bottom plot shows the reaching of the sliding surface and the rapid switching corresponds to the state trajectory reaching and remaining on the surface. The elevator deflection during this phase is commanded to go to zero therefore is not shown.

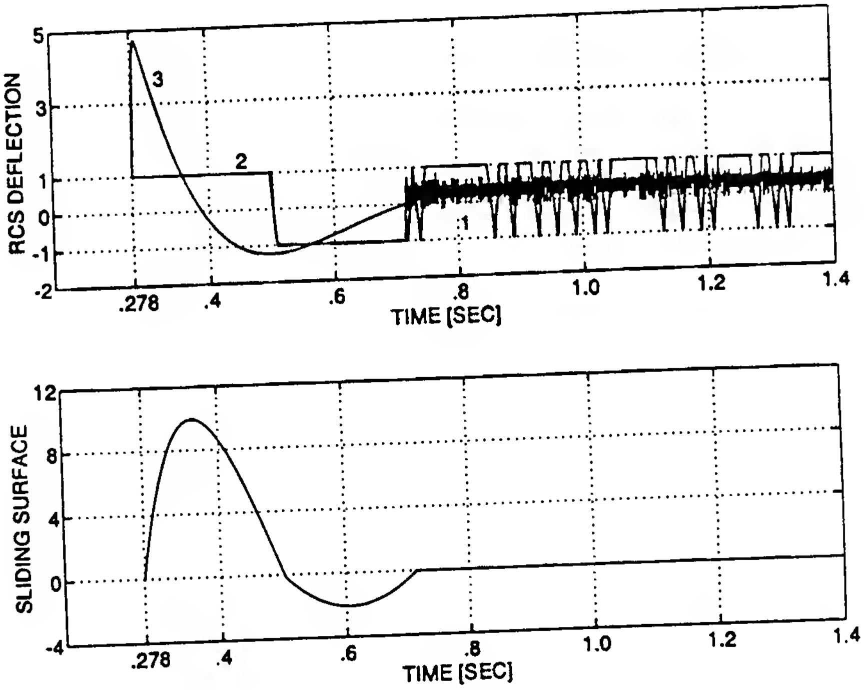


Figure 27. Control Activity and Sliding during Phase II

The Phase II trajectory is shown in figure 28. The solid line shows the trajectory followed by the center of mass. The missile's icon shows the attitude at various points. The velocity vector is shown by the vector lines, with the length of the vector being indicative of its magnitude.

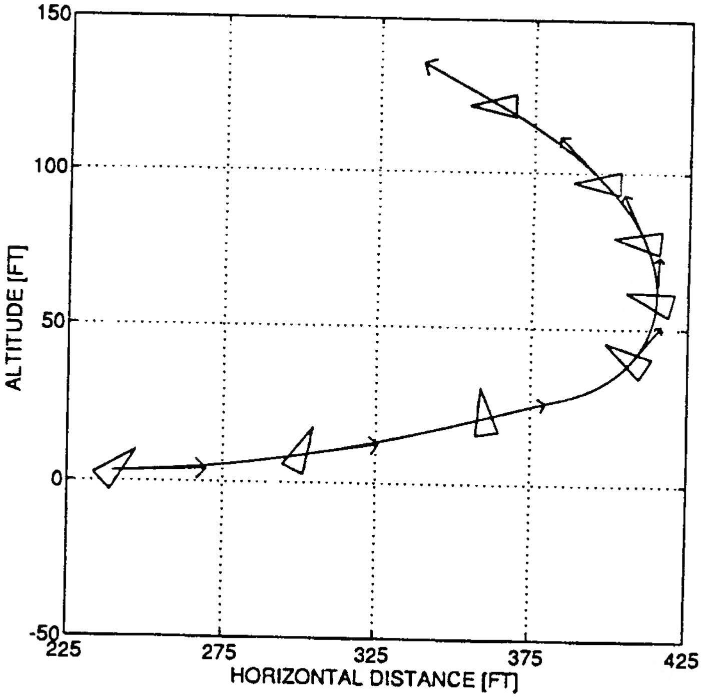


Figure 28. Trajectory during Phase II

The final phase is the acquistion of steady state conditions. The key variables are the pitch angle and the flight-path angle, which are required to reach a value of 180 degrees. This phase, like the previous one, uses a model following strategy to control the pitch attitude and the flight path, and two different models were used to achieve the final state, as described in section 4 of the report.

The first model contains the desired response for the states [$\theta$, q, $\gamma$]. The reason for having explicit $\theta$ and $\gamma$ was that we could drive these states to any desired final value. The second approach designs a model following controller based on the desired behavior of attitude and pitch rate only. The main purpose for choosing the two models is for comparison purposes, since this work is mainly a feasibility study. Phase III results are presented here for both the approaches.

With the first approach, the responses obtained are shown in Figure 29. The model response is indicated by the dotted line 2. The missile attitude and the flight path angle converge to the modelled responses. The pitch rate response on the other hand shows a high frequency oscillation of relatively small magnitude induced by the VSS operation. The average value of the pitch rate is the same as the modeled pitch rate.

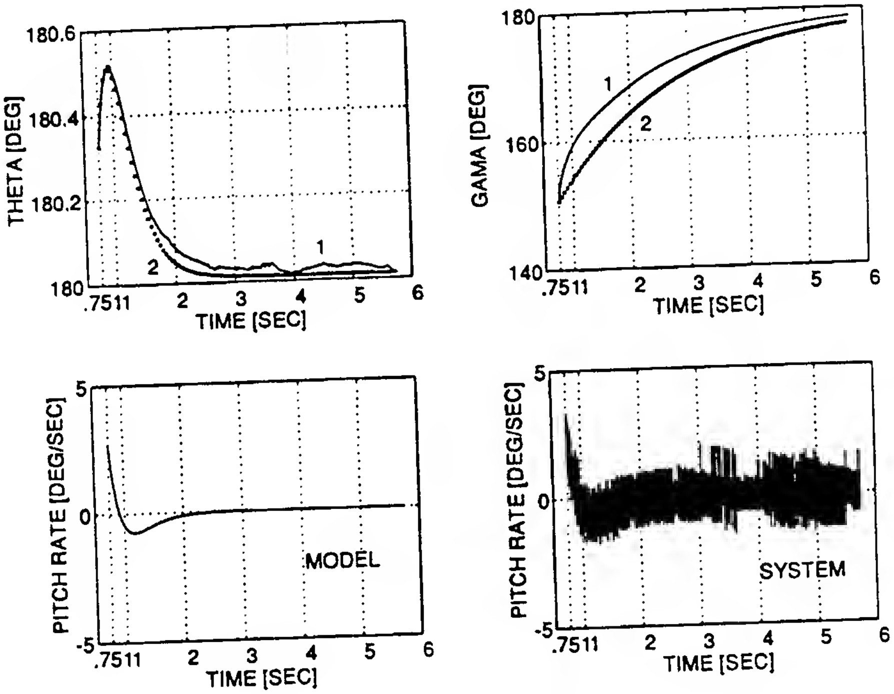


Figure 29. Angular Diplacement during Phase III

The elevator and reaction jet control activities are shown in Figure 30 and Figure 31 shows the corresponding sliding surfaces. The elevator response is a smooth function with no discontinuity, since the sliding surface 1 is not reached. On the other hand, the reaction jet thrusters show a high amount of chattering due to the system being in a sliding condition along surface 2.

The second approach uses two-degree of freedom model. The flight path angle appears as a disturbance in the equations describing the missile's attitude dynamics.

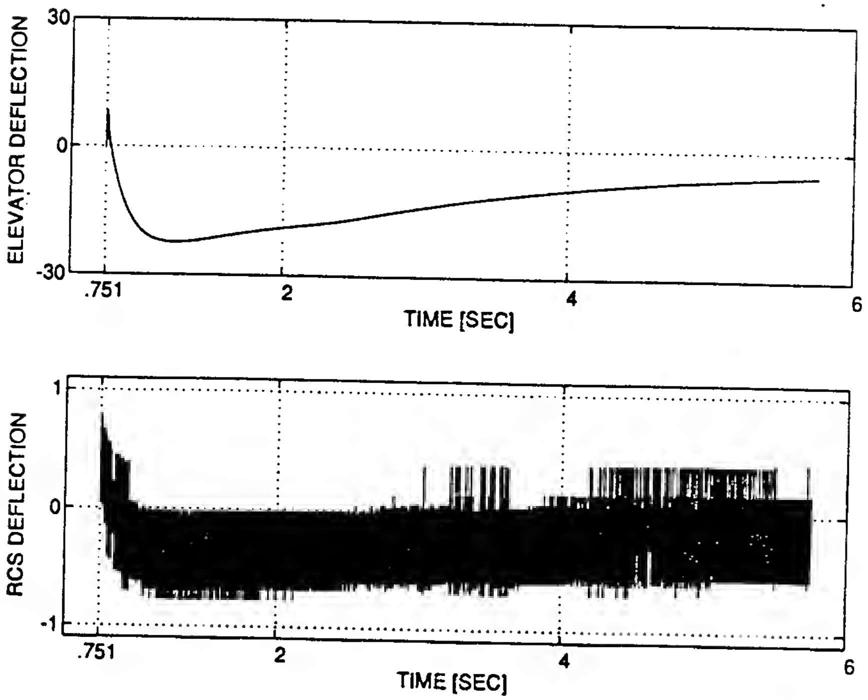


Figure 30. Control Activity during Phase III

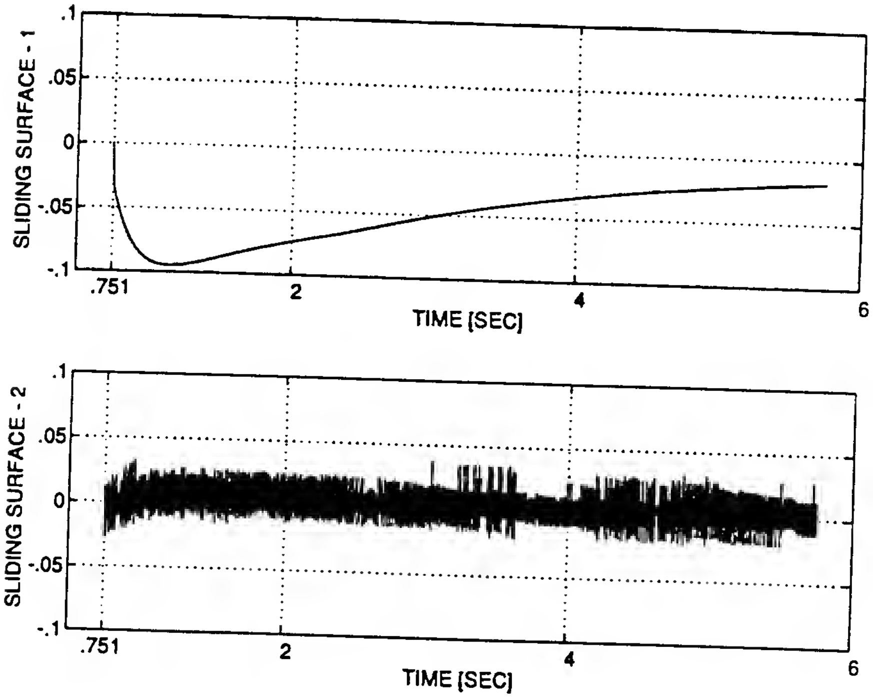


Figure 31. Sliding Surfaces during Phase III

The response for the model (line 2) and for the system's states are shown in Figure 32. The pitch rate responses for the model and the missile are shown separately in the second and third charts, with the vehicle's chattering characteristics, as before, linked to the VSS control structure. Figure 33 replicates the pitch rate behavior and the sliding surface over a shorter period of time, for clarity's sake.

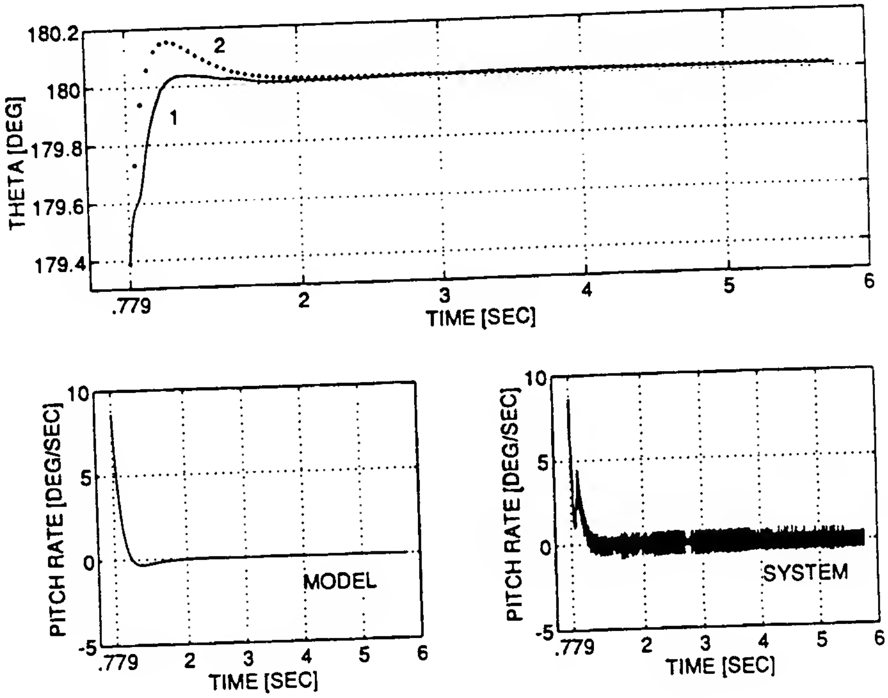


Figure 32. Motion Behavior during Phase III (Model Following of Attitude only)

Figure 34 shows the elevator and reaction jet control activities. The system that is being controlled is now a second order one, and two controls are available for this purpose. The control activity was broken down into two components, with perfect-model following being performed by the elevator, while the reaction jet's task was to take care of any modelling uncertainties and the disturbance introduced by the flight path angle. Latter results will show that this approach yields smaller turn radius and faster achievement of the steady state values.

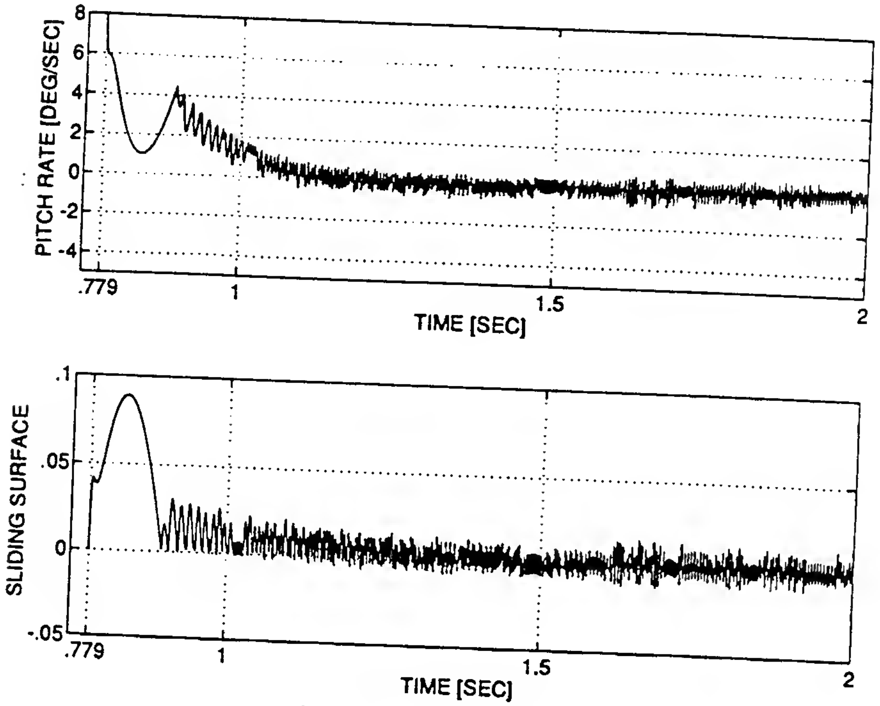


Figure 33. Pitch Rate and Sliding Surface (Model Following of Attitude only)

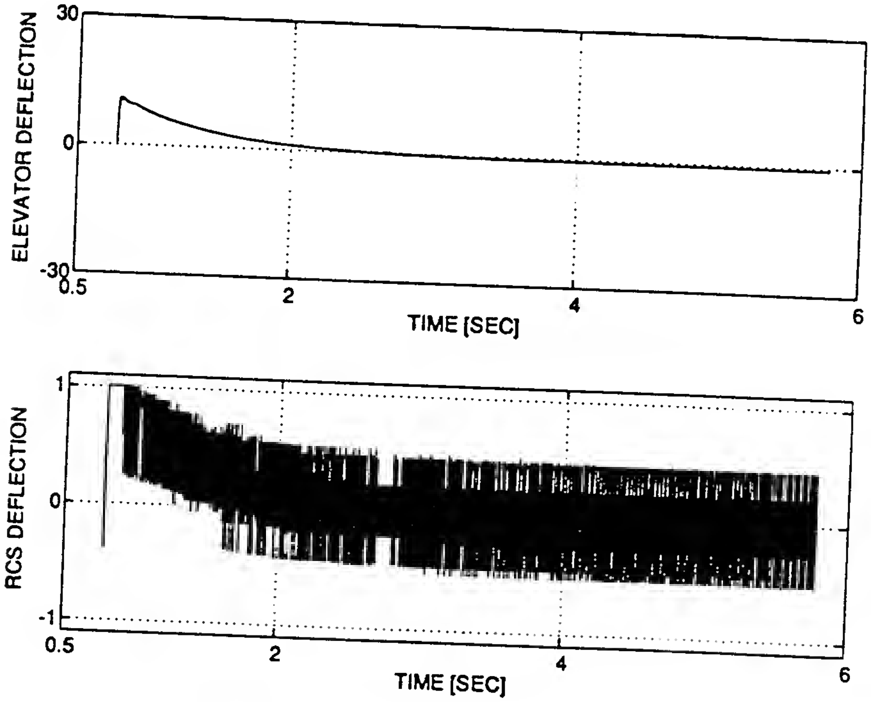


Figure 34. Control Activity during Phase III (Model Following of Attitude only)

The reaction jet's thrust does not reach the nominal value (corresponding to +/- 1 in Figure 34) because of the presence of actuator dynamics and deadband, which make the actuation not instantaneous as required by the VSS command.

Figures 35 and 36 compare the two model following approaches. Approach 2 produces a smaller turn radius as seen in Figure 35. The flight path angle, however, is nearly zero for both cases.

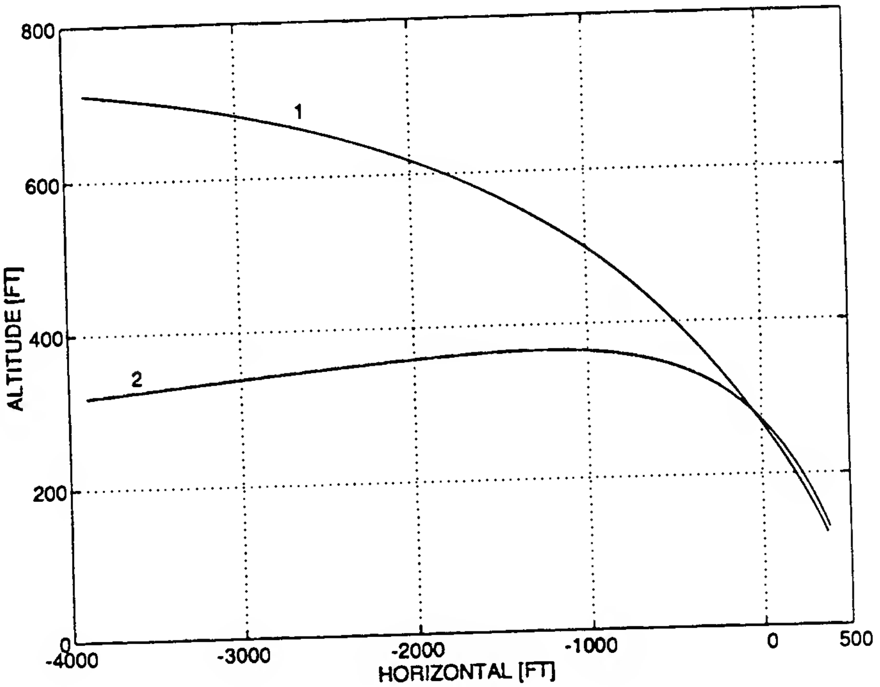


Figure 35. Trajectory Comparison for Phase III

Combining the three phases, yields the complete maneuver. Figure 37 shows the angles $\alpha$, $\theta$, and $\gamma$ only over the time period of one second, since the key characteristics belong to this time frame. The time grids show start and end of Phases I, II and III. The discontinuity in rate, during transition, is explained by the absence of any smoothing factor during the switching from one type of autopilot structure to another.

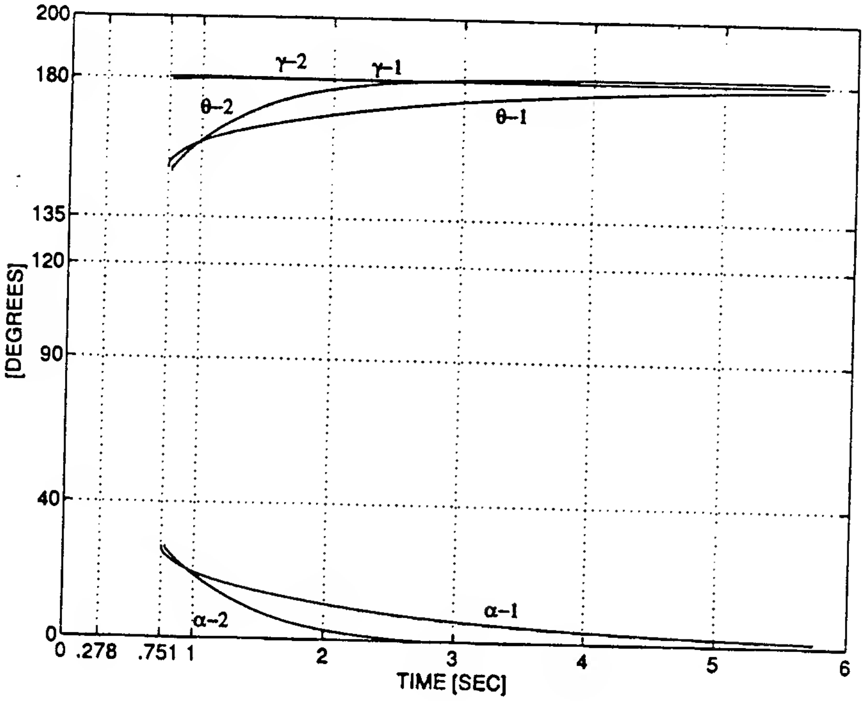


Figure 36. Angles Comparison for Phase III

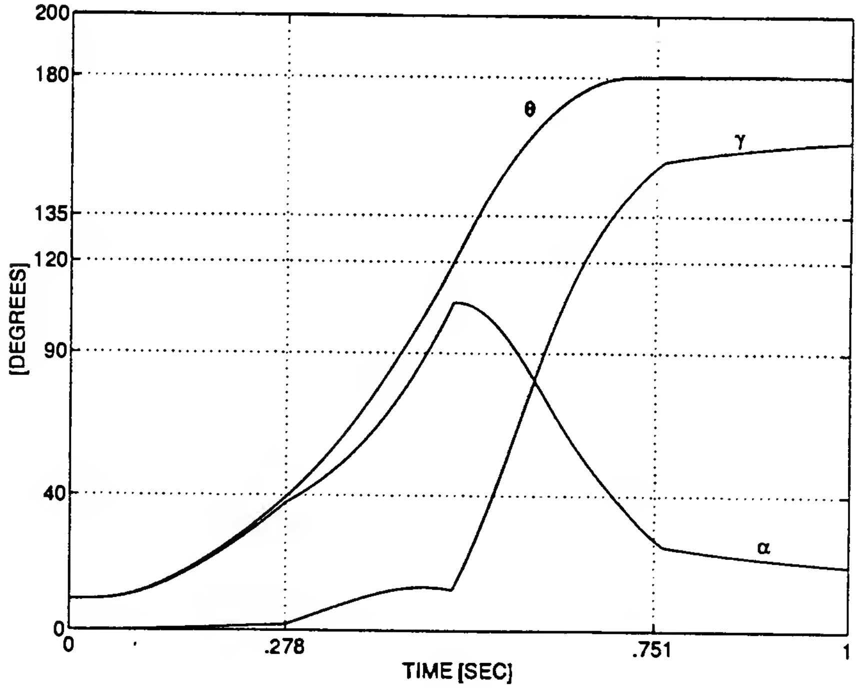


Figure 37. Angular Behavior during the Entire Maneuver

During Phase II the main engine is fired at a predefined attitude. The engine remain active until the speed reaches the final desired value. The range of activity can be seen in Figure 37 between 0.278 and 0.751 seconds with the discontinuities in $\alpha$ and $\gamma$.

Figure 38 illustrates the complete maneuver. The velocity vector is scaled with respect to the maximum speed. To achieve the correct perspective, the x and the y axes are drawn to the same size.

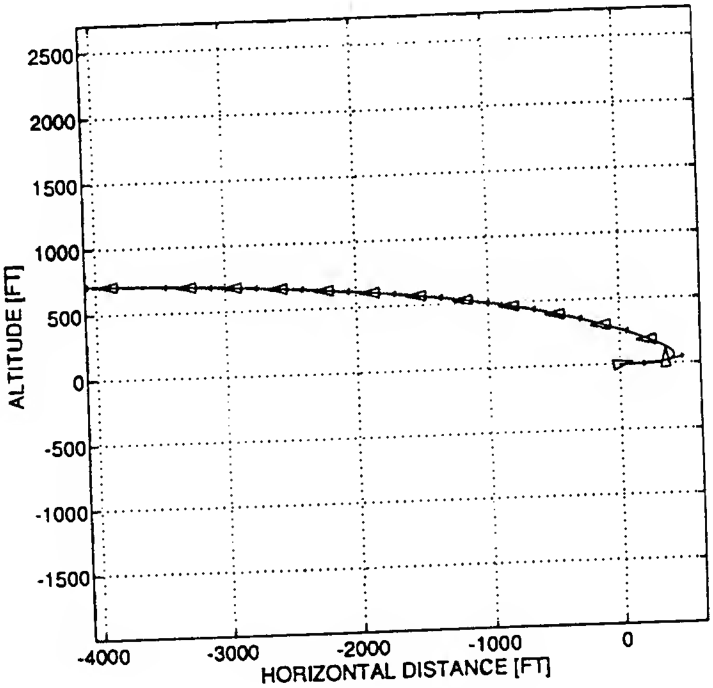


Figure 38. Vehicle's Trajectory during the Entire Maneuver

### 5.2.2. Sensitivity Analysis

A sensitivity and robustness analysis was carried out to show the effectiveness of the designed control system. A detailed description of time histories is presented in Appendix A3. In this section, some comments based on the resulting trajectories are presented, noting that all the curves were obtained by keeping the nominal gains with no gain scheduling. Also, both model following approaches for Phase III are included.

A parameter changed in the analysis was the reference speed. Simulations initiating the maneuver at Mach numbers of 0.6 and 1.2 were performed. Figure 39 shows the trajectory comparison using approach 1, and Figure 40 shows the trajectories using the second approach. During this simulation, the other parameters were kept constant and equal to their nominal value.

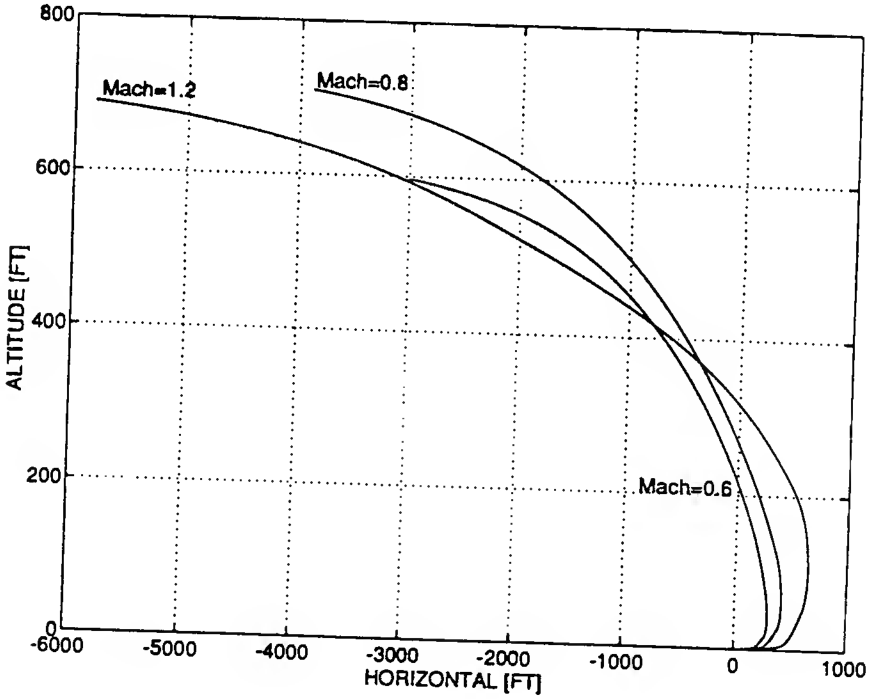


Figure 39. Trajectory Comparison at different Mach Number (Model Following of θ, q, and γ)

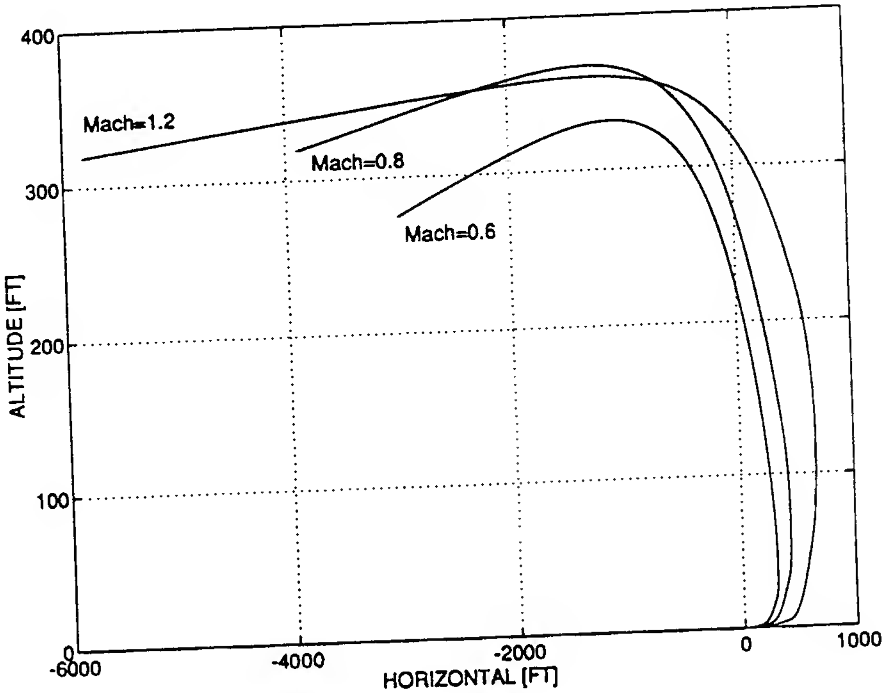


Figure 40. Trajectory Comparison at different Mach Number (Model Following of θ, q)

Figures 41 and 42 compare trajectories obtained by changing the attitude for initial firing of the main engine (120, 130 140 degrees), again according to the two model following approaches used for Phase III. The nominal speed corresponds to Mach 0.8, with nominal setting for the reaction jet's thrust, equal to 500 pounds.

Trajectory comparison for the nominal case, with two different reaction jet's thrust values (500 and 1000 pounds) is shown in Figures 43 and 44.

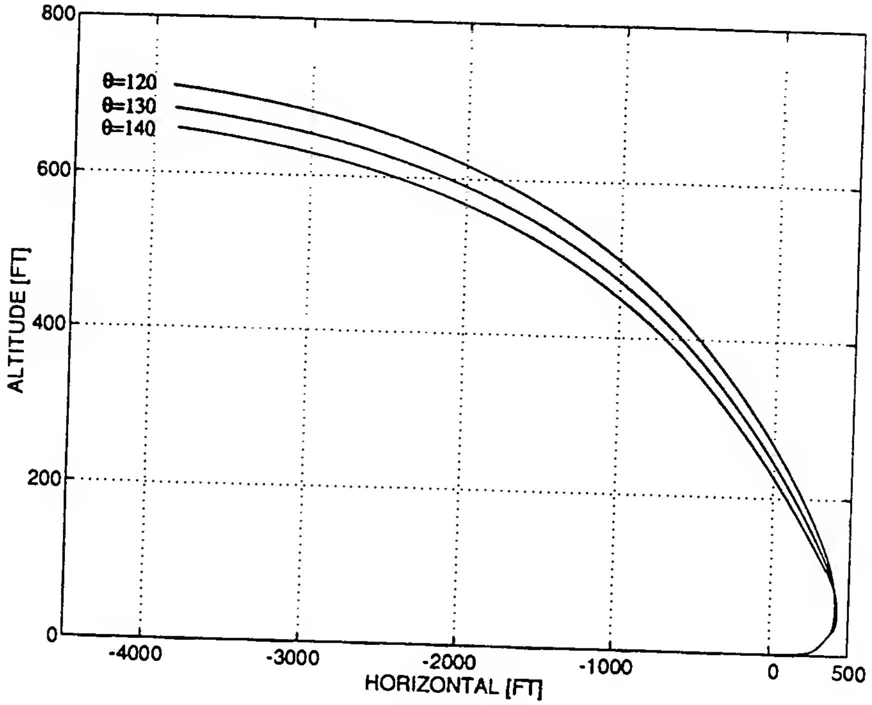


Figure 41. Trajectory Comparison at different Main Engine firing (Model Following of θ, q, and γ)

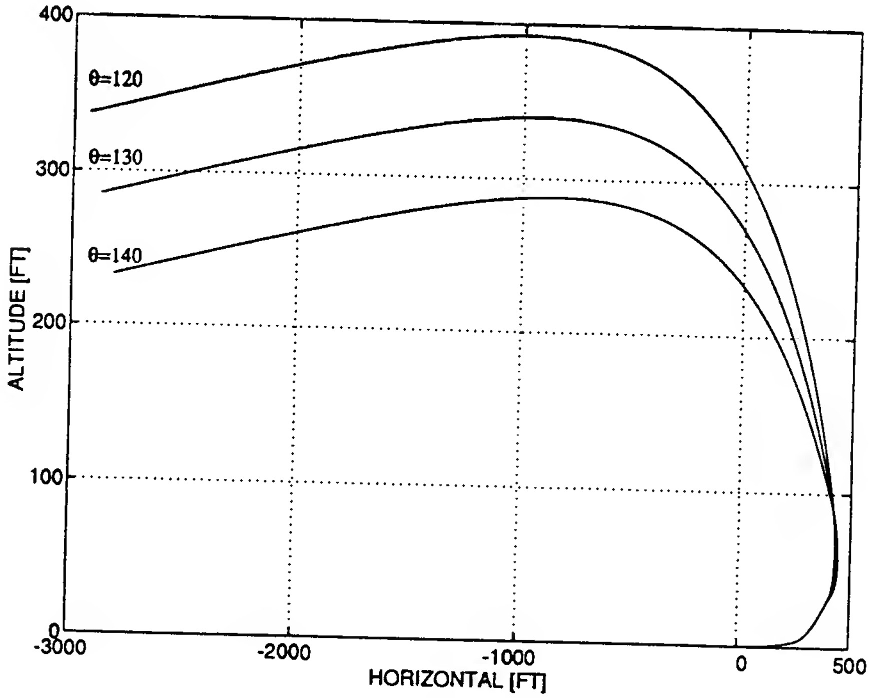


Figure 42. Trajectory Comparison at different Main Engine firing (Model Following of θ, q)

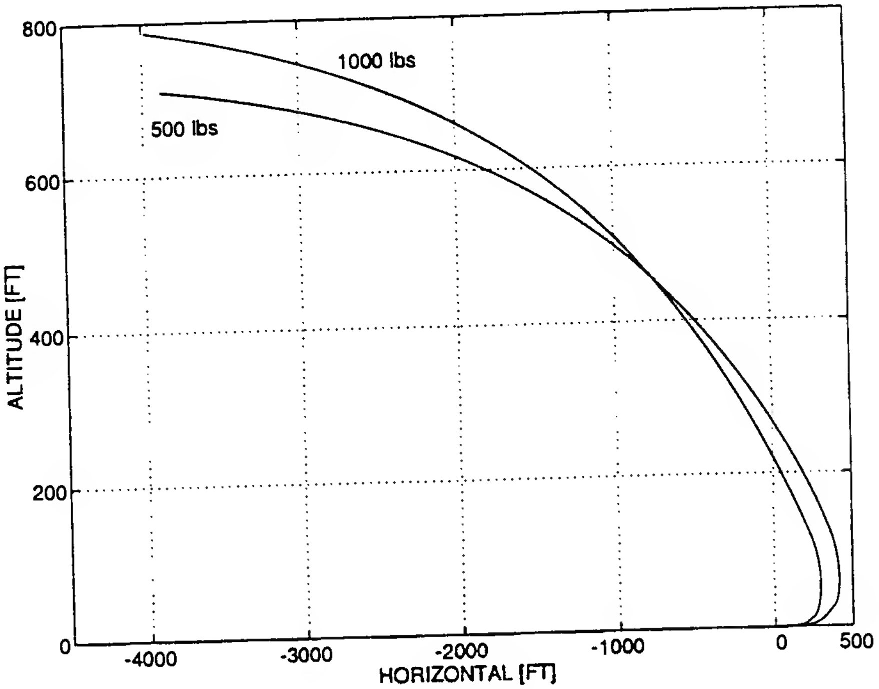


Figure 43. Trajectory Comparison at different RCS Thrust (Model Following of θ, q, and γ)

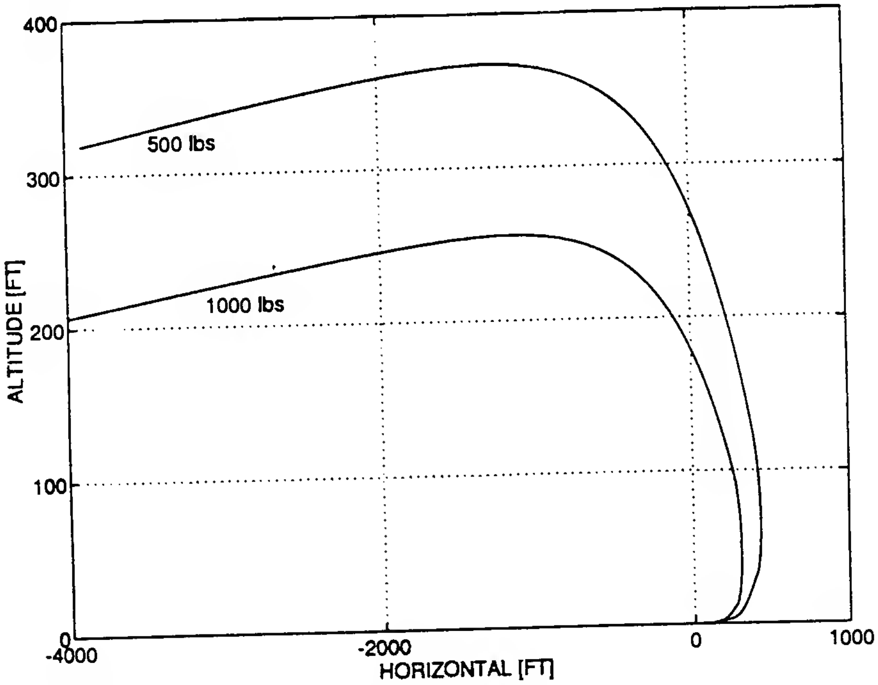


Figure 44. Trajectory Comparison at different RCS Thrust (Model Following of θ, q)

Finally a frequency response analysis was carried out. The problem being nonlinear in nature following approach was followed to obtain the systems frequency response. Phase I of the system was excited by an impulse input and system response was obtained. From the equally spaced time response of the system discrete frequency response was obtained by using Fast Fourier Transform. This was then interpreted into analog frequency, the key concern being the possibility of exciting the missile's bending modes. The first bending mode frequency depends on the diameter and weight (assuming no change in stiffness and material properties). For a 5 inch diameter and a 225 pound vehicle, the first bending mode frequency was found to be of the order of 30 hertz. Since the computed closed loop bandwidth is around 2 hertz, with large magnitude attenuation beyond this value, we concluded that flexibility of the structure is not an issue at this point.

Another aspect that was investigated, is the potential frequency aliasing due to sampling rates used for the digital implementation of the autopilot. A step size of 0.001 seconds was used in the simulations, leading to a Nyquist sampling frequency of about 100 hertz. Again, this is well over the bandwidth of the system.

### 5.3. Simulation Software

The computational aspects of the work, autopilot design and simulation were carried out using Matlab® version 4.1. The simulation code ACTMS (Alternate Control Technology Missile Simulation) was developed making extensive use of the block diagram capabilities of the available toolboxes and SIMULINK™. The computer used for the work was a SUN workstation, however the code will run on other platforms, provided version compatibility is satisfied. A description of the input/output properties of the mfiles developed for the simulation can be found in Appendix A2. A diskette with the source code is included in the report.

The input of physical parameters, state initialization are entered using an input file (**init_actms.m**), which defines missile parameters,speed, range of main engine operation, angle of attack range below stall, load factor command value, model desired values for attitude and flight path angles, and integration characteristics.

The simulation drivers are files **actms.m** and **SimAll.m**. They have two different versions according to the implementation of the model following relative to Phase III.

The three phases are defined by files **phaseI.m, phaseII.m**, and **phaseIII.m**, with the results of the simulation being written on data files **part4.mat, part5.mat, part6.mat**, and **part7.mat** (relative to the entire maneuver).

The models necessary to the simulation, inclusive of gain computation, are determined by the file **actdyn.m**. System matrices and gains are stored into datafiles **wrksp.mat, gain.mat**. Mfiles for plotting the results are also available (**mplot.m** and **mplot1.m**).

The file **aero.m** contains the lift and drag profiles in a matrix format, as functions of Mach number and angle of attack, as computed by DATCOM and/or derived analytically for the region above stall. Linear interpolation is used to extract values in-between. Of course different vehicles require appropriate aerodynamic data. The files responsible for adjusting the angle of attack value in order to read data beyond 90 degrees are **adjustalp.m** and **clause.m.**

The flowchart relative to the code is shown in Figure 45.

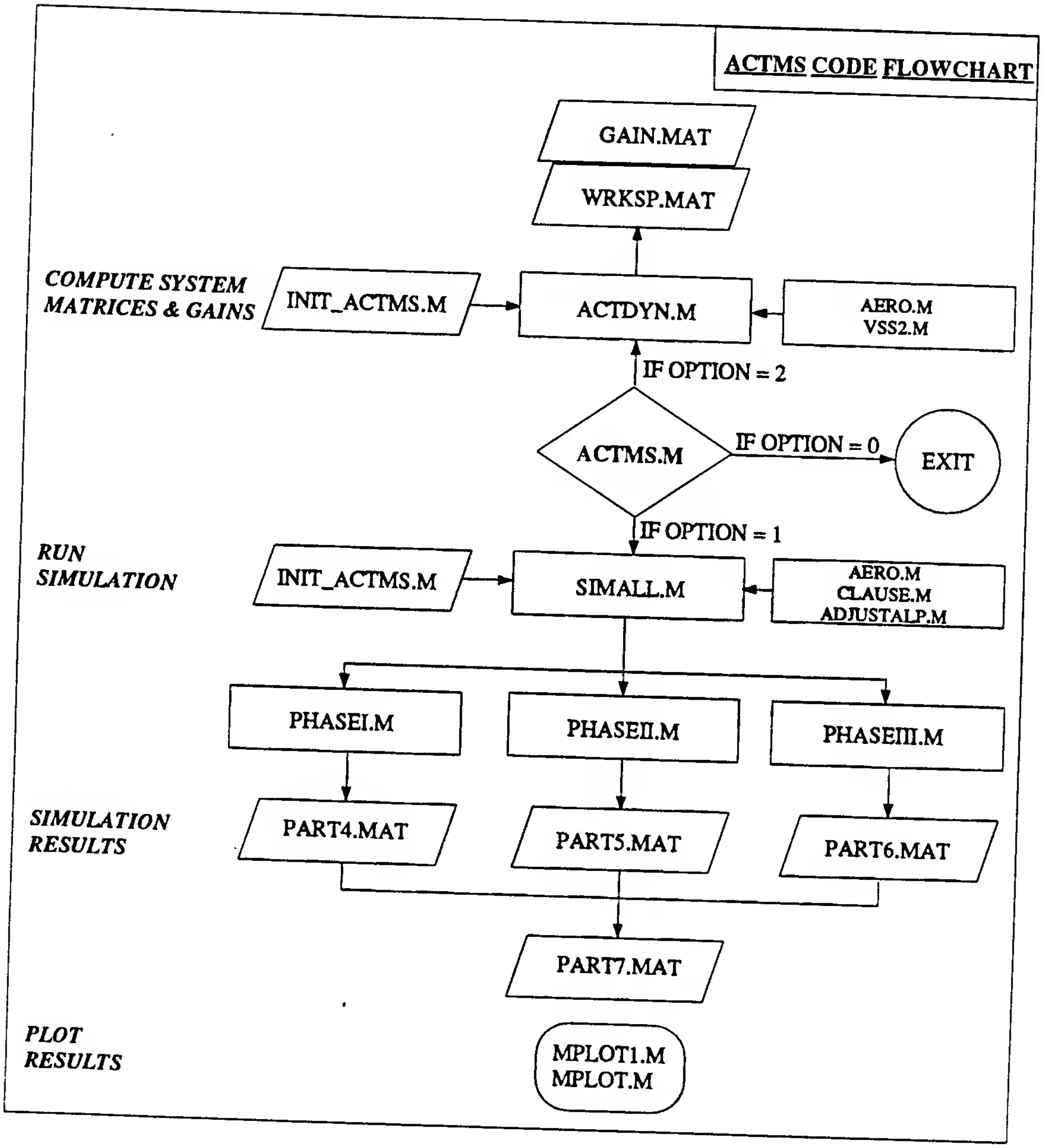


Figure 45. Simulation Code Flowchart

## 6. Conclusions and Recommendations

Conclusions

The following conclusions were reached at the end of work done under the present grant:

(1) Variable stucture control methods were applied with success to the feasibility design of a pitch autopilot, using a reduced size elevator and reaction jets as controllers. Different control structures were used during the chosen test maneuver.

(2) The autopilot did not require gain scheduling for a wide range of parameter variations.

(3) A simpler autopilot structure could be achieved by changing the control logic for Phase I, from a load factor command to one similar to that of the Phase II and Phase III.

Recommendations

The encouraging results of the present research warrant further work in several areas, in order to achieve a point design. These are:

(1) The study of optimal nonlinear trajectories, that will take full advantage of the added propulsive control capabilities. This study would definitely involve the definition of missile agility metrics, similar to those available for aircraft.

(2) The analysis of guidance laws capable of implementing acceleration and/or rate command required to achieve point (1).

(3) Autopilot design and control authority selection using output feedback VSS, since not all the states are used by the autopilot. This design would include analytical contributions in variable structure theory, as well as classical channel integration due to the three-dimensional aspects of the problem.

(4) Study of optimal reaction jet characteristics, such as amplitude and frequency modulated actuators to be used for pitch and yaw commands, as well as roll stabilization.

## 7. References

[1] Innocenti, M., Thukral, A., "Simultaneous Reaction Jet and Aerodynamic Control of Missile Systems", AIAA-93-3739 Guidance, Navigation and Control Conference, Monterey, California, August 1993.

[2] Innocenti, M., "Preliminary Missile Autopilot using Reaction Jet and Aerodynamic Control", Final Report RDL-33, AFOSR Summer Faculty Program, Wright Laboratory, Armament Directorate, Eglin AFB, August 1992.

[3] Weil, R.D., Wise, K.A., "Blended Aero & Reaction Jet Missile Autopilot Design using VSS Techniques", Proc. 30th IEEE CDC, Brighton, UK, December 1991.

[4] Jones, J., "Alternate Control Technology Program", WL/MNAV IRD Presentation, Wright Laboratory, Armament Directorate, Eglin AFB, August 1992.

[5] Emel'yanov, S.,V., "Design of Variable Structure Control Systems with Discontinuous Switching Functions", Engineering Cybernetics, 1, 1964.

[6] Utkin, V.I., Sliding Modes and Their Application in Variable Structure Systems, MIR, Moscow, 1978.

[7] Itkis, U., Control Systems of Variable Structure, Wiley, New York, 1976.

[8] Drazenovic, B., "The Invariance Condition in Variable Structure Systems", Automatica, Vol. 5, 1969, pp. 287-295.

[9] Zinober, A.S.I., Deterministic Control of uncertain Systems, IEE-40, Control Engineering Series, 1990.

[10] DeCarlo, R.A., Zak, S.H., Matthews, G.P., "Variable Structure Control of Nonlinear Multivariable Systems: a Tutorial", Proc. IEEE, Vol.76, No.3, March 1988.

[11] Calise, A.J., Kramer, F.S., "A Variable Structure Approach to Robust Control of VTOL Aircraft", AIAA. JGCD, Vol.7, 1984.

[12] Mudge, S.K., Patton, R.J., "Enhanced Assessment of Robustness for an Aircraft's Sliding Mode Controller", AIAA. JGCD, Vol.11, 1988.

[13] Slotine, J.J.E., Li, W., Applied Nonlinear Control, Prentice Hall, 1990.

[14] Vadali, S.R., "Variable Structure Control of Spacecraft Large Angle Maneuvers", AIAA. JGCD, Vol.9, No.2, 1986.

[15] Brierley, S.D., Longchamp, R., "Application of Sliding-Mode Control to Air-Air Interception Problem", IEEE-TR-AES, Vol.26, No.2, March 1990.

[16] Bradshaw, A., Counsell, J.M., "Design of Autopilots for High Performance Missiles", IMechE, Vol.206, 1992.

[17] Ryan, E.P., Corless, M., "Ultimate Boundedeness and Asymptotic Stability of a Class of Uncertain Dynamical Systems via Continuous and Discontinuous Feedback Control", IMA Journal, Vol.1, 1984.

[18] Ambrosino, G., Celentano, G., Garofalo, E., "Variable Structure Model Reference Adaptive Control Systems", Int. J. Contr., Vol.39, No.6, 1984.

[19] Balestrino, A., De Maria, G., Zinober, A.S.I., "Nonlinear Adaptive Model Following Control", Automatica, Vol.20, No. 5, 1984.

[20] Thukral, A., Innocenti, M., "Controls Design Challenge: A Variable Structure Approach," accepted for publication: AIAA JGCD, January 1993.

[21] Hung, J.Y., Gao, W., Hung, J.C., "Variable Structure Control: A Survey", IEEE-TR-IE, Vol.40, No.1, February 1993.

[22] Bruns, K.D., and others, "Missile DATCOM", WL-TR-91-3039.

[23] Hoerner, S., Practical Information on Fluid Dynamic Drag, Wiley 1958.

[24] Aerodynamics Data sheets.

[25] Innocenti, M., "Integrated Approach to the Guidance and Control of Aerospace Vehicles", Proposal, EOARD/AFOSR, November 1993.

[26] Blakelock, J.H., Automatic Control of Aircraft and Missiles, Wiley, 1991.

[27] Buhler, H., "Sliding Mode Control with Switching Command Devices", in Deterministic Control of Uncertain Systems, IEE-40, Control Engineering Series, 1990.

## Appendices

### A1. Gains

This Appendix contains the gain matrices used for the autopilot described in the report.

Phase I Autopilot

The gain matrices used in Eq. (48) are:

L_1 =[ -5.3517e-03  -1.1120e-02  -1.2727e-03  -9.5270e-01  -3.6759e-02
-4.9005e-03  -2.6377e-03  -2.6132e-03  -2.1420e-02  -9.8215e-01]

M_1 =[-1.3396e+00  -1.4005e-01  9.7002e-02  -6.8826e-01  2.3606e-01
5.0787e+00  2.2726e-01  1.5613e-01  -1.3907e+00  -5.8274e-01]

N_1 =[ -2.6758e-03  5.6661e-05  -3.8657e-04  3.0562e-03  1.2569e-04
-2.4502e-03  -2.5616e-04  1.7742e-04  -1.2589e-03  4.3177e-04]

The gain matrix G defining the sliding surface as defined on page 35 is given by:

G_1 =[-2.6792e+00  -2.8010e-01  1.9400e-01  -1.3765e+00  4.7212e-01
1.0157e+01  4.5452e-01  3.1226e-01  -2.7815e+00  -1.1655e+00]

Phase II Autopilot

The gain matrices in Eq. (54) are:

$K_1$_2 = [-2.8664e+00 -3.8649e-01]

$K_2$_2 = 2.8664e+00

$K_3$_2 = -1.6104e-01

The gain matrix G defining the sliding surface from Eq. (55) is:

G_2 = [7  1]

### Phase III Autopilot (method 1)

The model to be followed has a third order dynamics. The gain matrix $K_1$ in Eq. (61) is:

$K_1$_3 =[-4.9665e-01 5.0213e-02 -5.7851e-01
-4.4346e-01 -5.5621e-02 -1.3817e+00]

The gain matrices in Eq. (63) are:

L_3 =[ 8.7872e-02 3.1383e-02 -5.8779e+00
-9.7337e-02 -3.4763e-02 -8.5697e+00]

M_3 =[ 3.3804e-03 3.3804e-04 1.6667e-01
-1.6667e+00 -1.6667e-01 3.3804e-04]

N_3 =[ 1.0461e-02 1.0461e-03 -3.9186e-01
-1.1588e-02 -1.1588e-03 -5.7131e-01]

The gain matrix G defining the sliding surface is:

G_3 =[ 2.0283e-02 2.0283e-03 1.0000e+00
-1.0000e+01 -1.0000e+00 2.0283e-03]

### Phase III Autopilot (method 2)

The model to be followed has a second order dynamics. The gains in Eq. (66) are given by:

$K_x$_3 = [ 3.2843e-02 1.3254e-01]

$K_{um}$_3 = -4.0205e-01

$K_\gamma$_3 = -3.6921e-01

The gain G defining the sliding surface is in this case:

G_3 = [7 1]

## A2. Software Description

A short description of some of the more important mfiles is listed below. The file names and the function names are in italics.

*actms*

| | | |
|---|---|---|
| Mfile | *actms.m* | |
| Function | *actms* | |
| Input files | *Input/wrksp.mat* | System parameters are loaded |
| | *Input/gain.mat* | Model matrices and gain matrices for controls |
| Output files | None | |
| Functions | *SimAll.m* | Function simulates all the phases |
| | *actdyn.m* | Computes system matrices, model system matrices and gains |
| Messages | ACTMS Options:<br>(0) Exit<br>(1) Run Simulation<br>(2) Compute System Models/Gains<br>Selected Option ==><br>... SIMULATION OVER ...<br>... SYSTEM/MODELS/GAINS COMPUTED ... | |
| Error Messages | .. Incorrect entry .. | |

Description

*actms* (ALTERNATE CONTROL TECHNOLOGY MISSILE SIMULATION) is the driver routine making calls to *SimAll* and *actdyn* functions. *actdyn* fuction is called if some system parameter has been changed or new sets of gains are to be computed.

There is an error message if option number selected is an incorrect one.

*SimAll*

| | | |
|---|---|---|
| Mfile | *SimAll.m* | |
| Function | *SimAll* | |
| Input Files | *init_actms.m* | Initializing parameters, initial onditions |
| Output Files | *Output/part4.mat* | Saves Phase I variables t, y, yinert, u, unl, s |
| | *Output/part5.mat* | Saves Phase II variables t, y, yinert, u, unl, s, ym_2 |
| | *Output/part6.mat* | Saves Phase III variables t, y, yinert, u, unl, s, ym_3 |
| | *Output/part7.mat* | Saves for the entire maneuver, variables t, y, yinert, u, unl, s |
| Function Files | *ode23.m* | Matlab function for solving set of differential equations. |
| | *phaseI.m* | Simulink file for Phase I |
| | *phaseII.m* | Simulink file for Phase II |
| | *phaseIII.m* | Simulink file for Phase III |
| | *mplot.m* | Plots variables Nz, q, $\alpha$, $\gamma$, $\theta$, trajectory |
| | *mplot1.m* | Plots variables Speed, X, Z, $\delta$, $u_T$ |

Messages

Input filename (within quotes) ==>
Part 1 Done
Part 2 Done
Part 3 done
Ready for PHASE I simulation
Control C to press return or enter return ...
Running PHASE I ...
Plotting Phase I results
Saving Phase I results to part4.mat
Part 4 done (PHASE I)
Ready for PHASE II simulation
Control C to further simulation or else enter return ...
Running PHASE II ...
Plotting Phase II results
Saving Phase II results to part5.mat
Part 5 done (PHASE II)
Ready for PHASE III simulation
Control C to stop further simulation or else enter return ...
Running PHASE III ...
Plotting Phase III results
Saving Phase III results to part6.mat
Part 6 done (PHASE III)
Plotting entire simulation results to Figures 7, 8 ...
Saving entire simulation results to part7.mat
Part 7 Done.

Error Messages None

Description

Simulation of all the phases is carried out by this function. The three phases are formulated in terms of simple blocks. These blocks are then simulated using one of the MATLAB's routines for integrating or solving differential equations. Fuction *ode23* is called for simulating the phases.

There is almost a running commentary on the simulation status. Key message is the Control C .. message. The message tells the user that the MATLAB is ready for simulating a particular phase and that at this point the user has an option to Exit from *SimAll* and use SIMULINK capability to run the particular phase. This is helpful if the user is interested in performing parametric studies or optimizing the gains. The simulink block diagram for the phases can be simply called by typing in the corresponding mfile *phaseI*, *phaseII*, or *phaseIII*. To simulate the block diagram select option Start from the pull down Simulation menu.

Simulation results for each of the phases are plotted using mfiles *mplot.m* and *mplot1.m*. All the phases use same variable names to store the simulation data. After completion of each phase the data is therefore saved to separate mat files in sub-directory *Output*. The variables are:

| | |
|---|---|
| t | time |
| y | $[\gamma\ \theta\ \alpha\ A_z\ q\ \delta\ u_T]$ |
| yinert | $[U\ X\ W\ Z\ V\ d\gamma/dt]$ |
| ym_2 | $[\theta_m\ q_m]$ |
| ym_3 | $[\theta_m\ q_m\ \gamma_m]$ |
| unl | $[\delta_{nl}\ u_{Tnl}]$ |
| ulin | $[\delta_{lin}\ u_{Tlin}]$ |
| s | $[s_\delta\ s_{uT}]$ |

For each time instant, t, the variables are saved as a row of data. The angular units are in degrees. The state vector, $x=[\int A_z\ A_z\ q\ \delta\ u_T\ ]$ is not saved but is in the MATLAB's workspace area. The state vector has angular units in radians.

*actdyn*

| | | |
|---|---|---|
| Mfile | *actdyn.m* | |
| Function | *actdyn* | |
| Input Files | *init_actms.m* | |
| Output Files | *Input/wrksp.mat*<br>*Input/gain.mat* | |
| Function | *vss2.m* | Synthesis of gains for VSS |
| Messages | Appropriate messages | leading to the setting up and synthes of gains |
| Error Messages | | None |

Description

There are four basic parts to this routine. First part is related to finding the system matrices based on the aerodynamic data, speed, reaction thrust data. These are set in *init_actms.m*. The next three parts are for obtaining gains and models for the phases: Phase I, Phase II and Phase III. The models are defined in this file and if a new model values are required to be set then one has to change the values in this file.

The system matrices are saved in *Input/wrksp. mat* are: Asys_1 Bsys_1 Csys_1 Dsys_1 Asys_2 Bsys_2 Brcs_2 Asys_3 Bsys_3 Aelev Belev Celev Delev Atrcs Btrcs Ctrcs Dtrcs aaug baug caug daug.

The gains and models used for the simulation are saved in *Input/gain.mat* : L_1 M_1 N_1 G_1 Q_1 la3st_1 rhode_1 rhot_1 Am_2 Bm_2 Cm_2 Dm_2 Kx_2 Kum_2 Ksys_2 G_2 Am_3 Bm_3 Cm_3 Dm_3 L_3 M_3 N_3 G_3 Q_3 la3st_3 Kx_3 Kum_3 rhode_3 rhot_3 *phaseI, phaseII, phaseIII*

| | | |
|---|---|---|
| Mfile | *phaseI.m, phaseII.m, phaseIII.m* (SIMULINK files) | |
| Input files | None | |
| Output files | None | |
| Functions | *clause.m* | Checks the range condition |
| | *adjustalp.m* | Aero data for $\alpha$ beyond 90 degree is obtained by adjusting the value of $\alpha$. |
| Messages | | None |
| Error Messages | | None |

Description

These are block diagrams for Phase I, Phase II and Phase III respectively. Once all the variables are loaded into Matlab's workspace a phase can be simulated by clicking on <u>Simulation</u> option. This pops up a pull down menu. Select <u>Start</u> option from this and the simulation starts. If some variable is not defined an error message saying variable. The simulation results are stored into workspace. The variables are:

| | |
|---|---|
| t | time |
| y | $[\gamma\ \theta\ \alpha\ A_z\ q\ \delta\ u_T]$ |
| yinert | $[U\ X\ W\ Z\ V\ d\gamma/dt]$ |
| ym_2 | $[\theta_m\ q_m]$ |
| ym_3 | $[\theta_m\ q_m\ \gamma_m]$ |
| unl | $[\delta_{nl}\ u_{Tnl}]$ |
| ulin | $[\delta_{lin}\ u_{Tlin}]$ |
| s | $[s_\delta\ s_{uT}]$ |

For each time instant, t, the variables are saved as a row of data. The angular units are in degrees. The state vector, $x=[\int A_z\ A_z\ q\ \delta\ u_T\ ]$ is not saved but is in the MATLAB's workspace area. The state vector has angular units in radians.

## A3. Parametric Analysis

Simulation results for various flight conditions and parameter values are included here for purposes of reference and validation. The axes plots were all made using the same scale so that the results may be easily compared. Results relative to Phases I, II, and III are shown in Figures A-1 through A-26. Phase I and Phase II simulations are presented in Figures A-1 through A-8. The time histories shown in Figures A-9 through A-17 were generated using the 3-DOF model for Phase III control. Those presented in Figures A-18 through A-26 were obtained using the 2-DOF model for the final phase.

The parametric analysis values are described in Table A-1. Table A-2 shows a summary relationship between figure number, variables presented, and appropriate maneuver phase.

Finally, Figures A-27 and A-28 show the $C_L$, $C_D$ versus angle of attack curves at Mach numbers 0.6 and 1.2 respectively.

Table A-1

| RUN<br>Nos | Inital<br>Mach | Final<br>Mach | RCS<br>Thrust | θ<br>Engine ON |
|---|---|---|---|---|
| 1 | 0.6 | 0.6 | 500 | 120 |
| 2 | 0.6 | 0.6 | 1000 | 120 |
| 3* | 0.8 | 0.8 | 500 | 120 |
| 4 | 0.8 | 0.8 | 500 | 130 |
| 5 | 0.8 | 0.8 | 500 | 140 |
| 6 | 0.8 | 0.8 | 1000 | 120 |
| 7 | 1.2 | 1.2 | 500 | 120 |
| 8 | 1.2 | 1.2 | 1000 | 120 |

Table A-2

| Figure Nos. | Output Plotted | Corresponding Phase |
|---|---|---|
| Figure A-1 | $N_z$, $q$ | Phase I |
| Figure A-2 | $\theta, \alpha, \gamma$ | Phase I |
| Figure A-3 | $\delta, u_T$ | Phase I |
| Figure A-4 | $-Z$ vs $X$ | Phase I |
| Figure A-5 | $N_z$, $q$ | Phase II |
| Figure A-6 | $\theta, \alpha, \gamma$ | Phase II |
| Figure A-7 | $\delta, u_T$ | Phase II |
| Figure A-8 | $-Z$ vs $X$ | Phase II |
| Figure A-9 | $N_z$, $q$ | Phase III |
| Figure A-10 | $\theta, \alpha, \gamma$ | Phase III |
| Figure A-11 | $\delta, u_T$ | Phase III |
| Figure A-12 | $-Z$ vs $X$ | Phase III |
| Figure A-13 | $N_z$, $q$ | Complete Maneuver |
| Figure A-14 | $\theta, \alpha, \gamma$ | Complete Maneuver |
| Figure A-15 | $\delta, u_T$ | Complete Maneuver |
| Figure A-16 | $-Z$ vs $X$ | Complete Maneuver |
| Figure A-17 | Speed | Complete Maneuver |

[Using 2-DOF for the final phase control. Phase I and Phase II are identical to previous plots Figures A-1 through A-8]

| | | |
|---|---|---|
| Figure A-18 | $N_z$, $q$ | Phase III |
| Figure A-19 | $\theta, \alpha, \gamma$ | Phase III |
| Figure A-20 | $\delta, u_T$ | Phase III |
| Figure A-21 | $-Z$ vs $X$ | Phase III |
| Figure A-22 | $N_z$, $q$ | Complete Maneuver |
| Figure A-23 | $\theta, \alpha, \gamma$ | Complete Maneuver |
| Figure A-24 | $\delta, u_T$ | Complete Maneuver |
| Figure A-25 | $-Z$ vs $X$ | Complete Maneuver |
| Figure A-26 | Speed | Complete Maneuver |

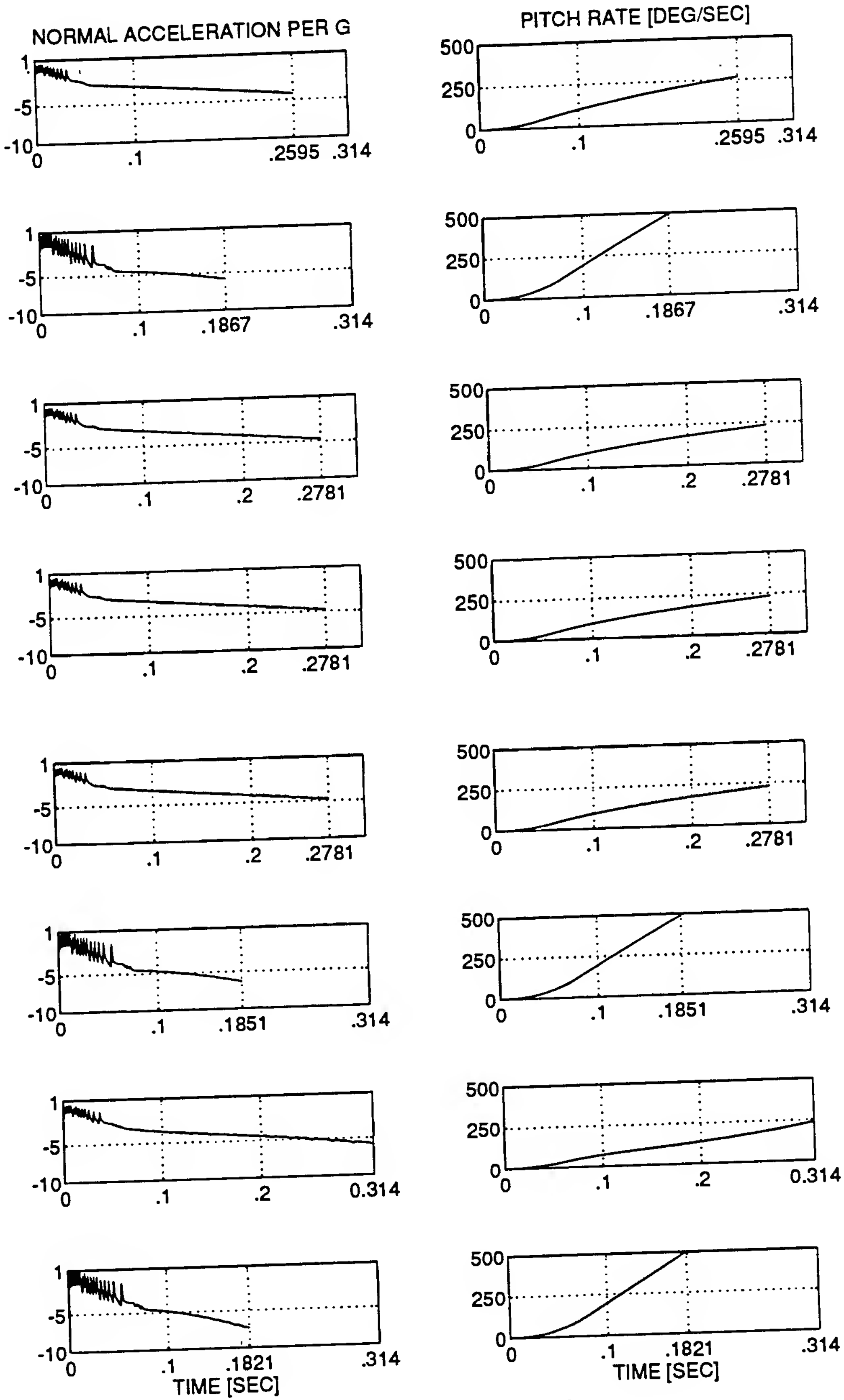

NORMAL ACCELERATION PER G
PITCH RATE [DEG/SEC]
1
-5
-10
0
.1
.2
.2595
.314
.1867
.2781
.1851
0.314
.1821
500
250
0
TIME [SEC]
TIME [SEC]

Figure A-1

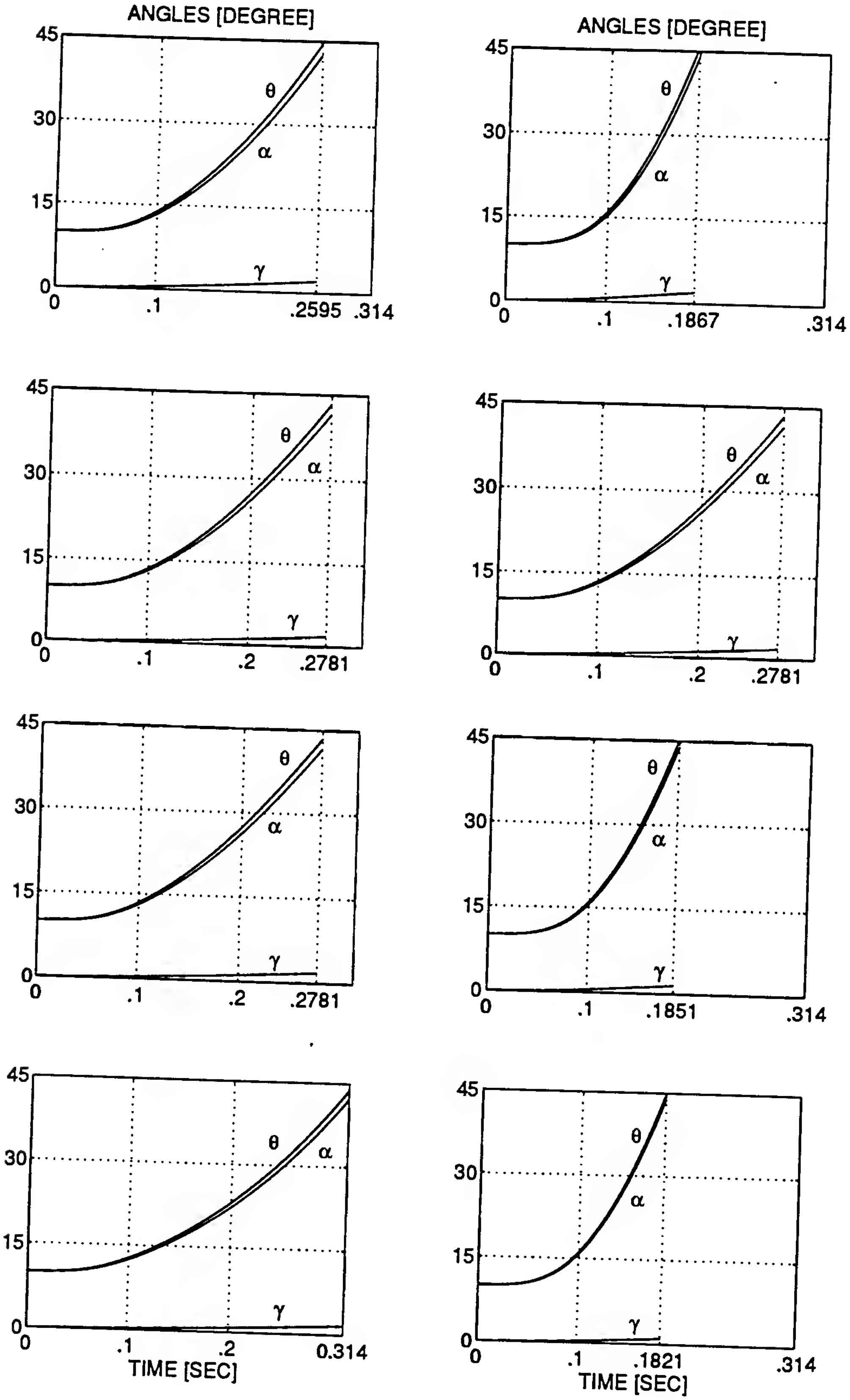

ANGLES [DEGREE]
45
30
15
0
θ
α
γ
0
.1
.2595
.314
ANGLES [DEGREE]
45
30
15
0
θ
α
γ
0
.1
.1867
.314
45
30
15
0
θ
α
γ
0
.1
.2
.2781
45
30
15
0
θ
α
γ
0
.1
.2
.2781
45
30
15
0
θ
α
γ
0
.1
.2
.2781
45
30
15
0
θ
α
γ
0
.1
.1851
.314
45
30
15
0
θ
α
γ
0
.1
.2
0.314
TIME [SEC]
45
30
15
0
θ
α
γ
0
.1
.1821
.314
TIME [SEC]

Figure A-2

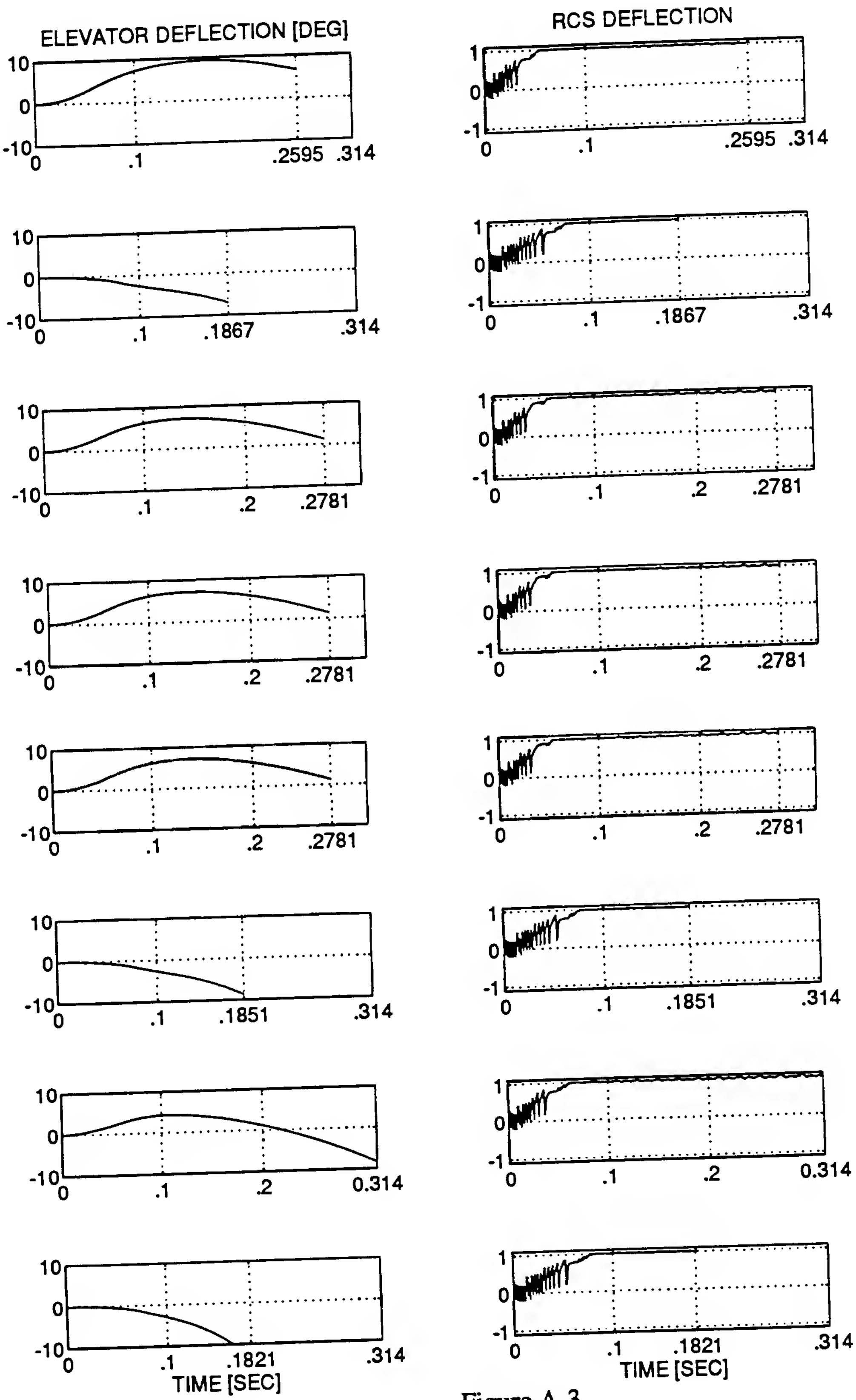

ELEVATOR DEFLECTION [DEG]
RCS DEFLECTION
10
0
-10
1
0
-1
0
.1
.2
.1867
.1851
.1821
.2595
.2781
.314
0.314
TIME [SEC]
TIME [SEC]

Figure A-3

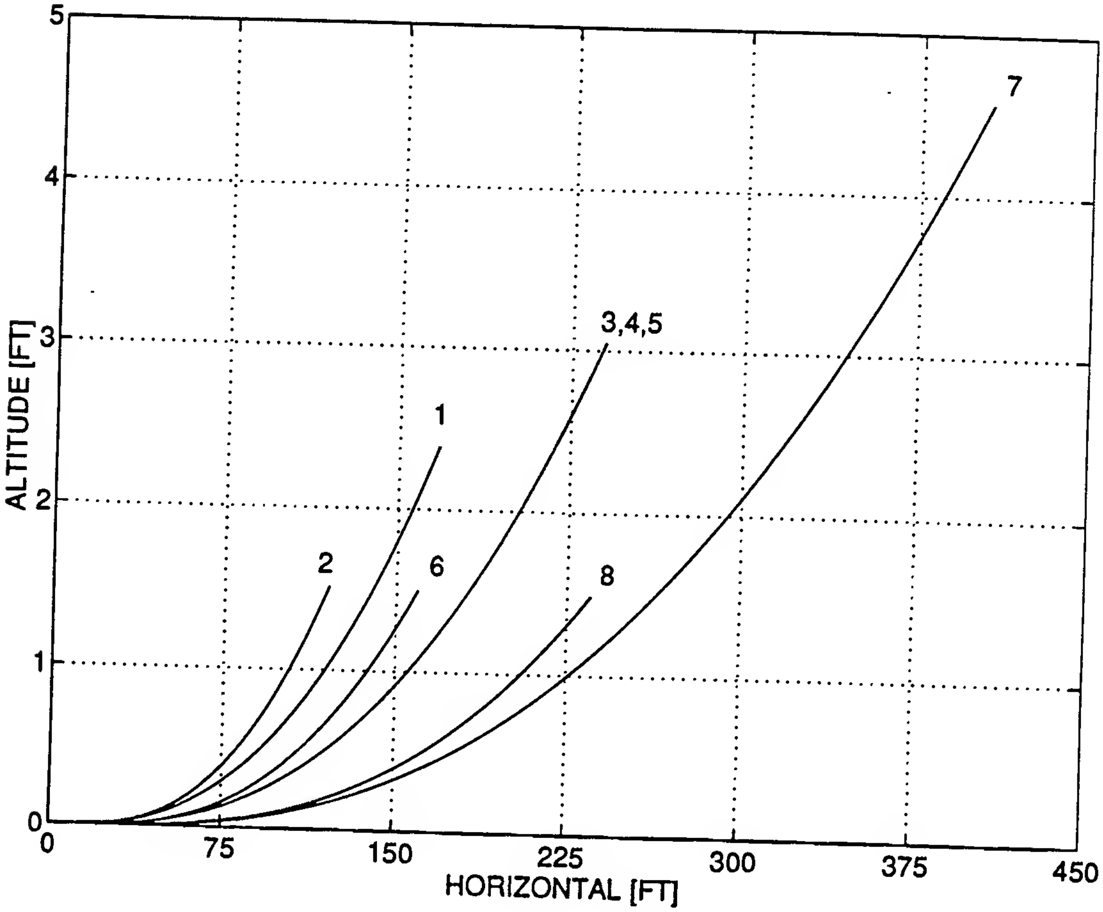

5
4
3
2
1
0
ALTITUDE [FT]
0
75
150
225
300
375
450
HORIZONTAL [FT]
7
3,4,5
1
2
6
8

Figure A-4

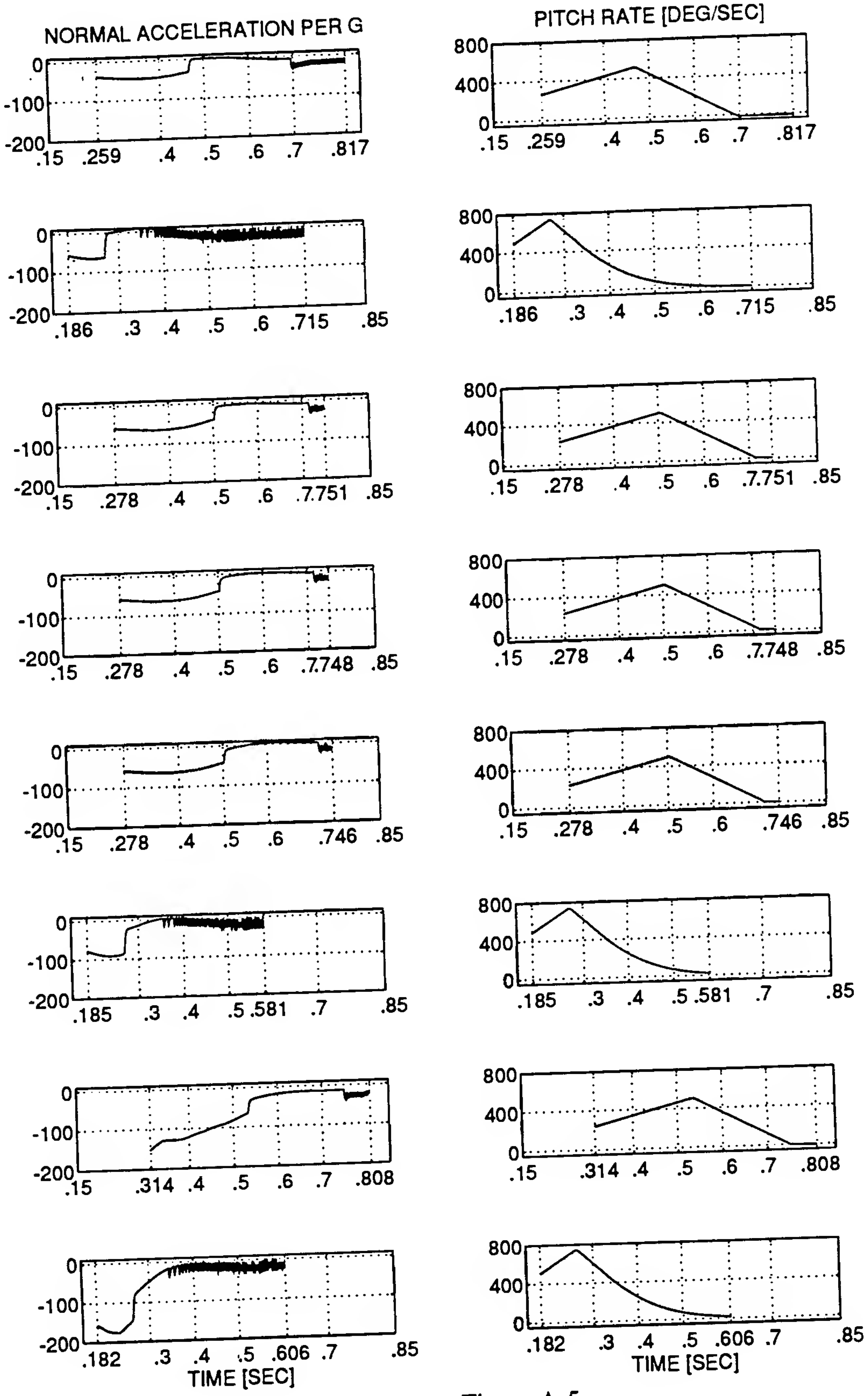

NORMAL ACCELERATION PER G
PITCH RATE [DEG/SEC]
0
-100
-200
.15 .259 .4 .5 .6 .7 .817
800
400
0
.15 .259 .4 .5 .6 .7 .817
.186 .3 .4 .5 .6 .715 .85
.186 .3 .4 .5 .6 .715 .85
.15 .278 .4 .5 .6 .7.751 .85
.15 .278 .4 .5 .6 .7.751 .85
.15 .278 .4 .5 .6 .7.748 .85
.15 .278 .4 .5 .6 .7.748 .85
.15 .278 .4 .5 .6 .746 .85
.15 .278 .4 .5 .6 .746 .85
.185 .3 .4 .5 .581 .7 .85
.185 .3 .4 .5 .581 .7 .85
.15 .314 .4 .5 .6 .7 .808
.15 .314 .4 .5 .6 .7 .808
.182 .3 .4 .5 .606 .7 .85
.182 .3 .4 .5 .606 .7 .85
TIME [SEC]
TIME [SEC]

Figure A-5

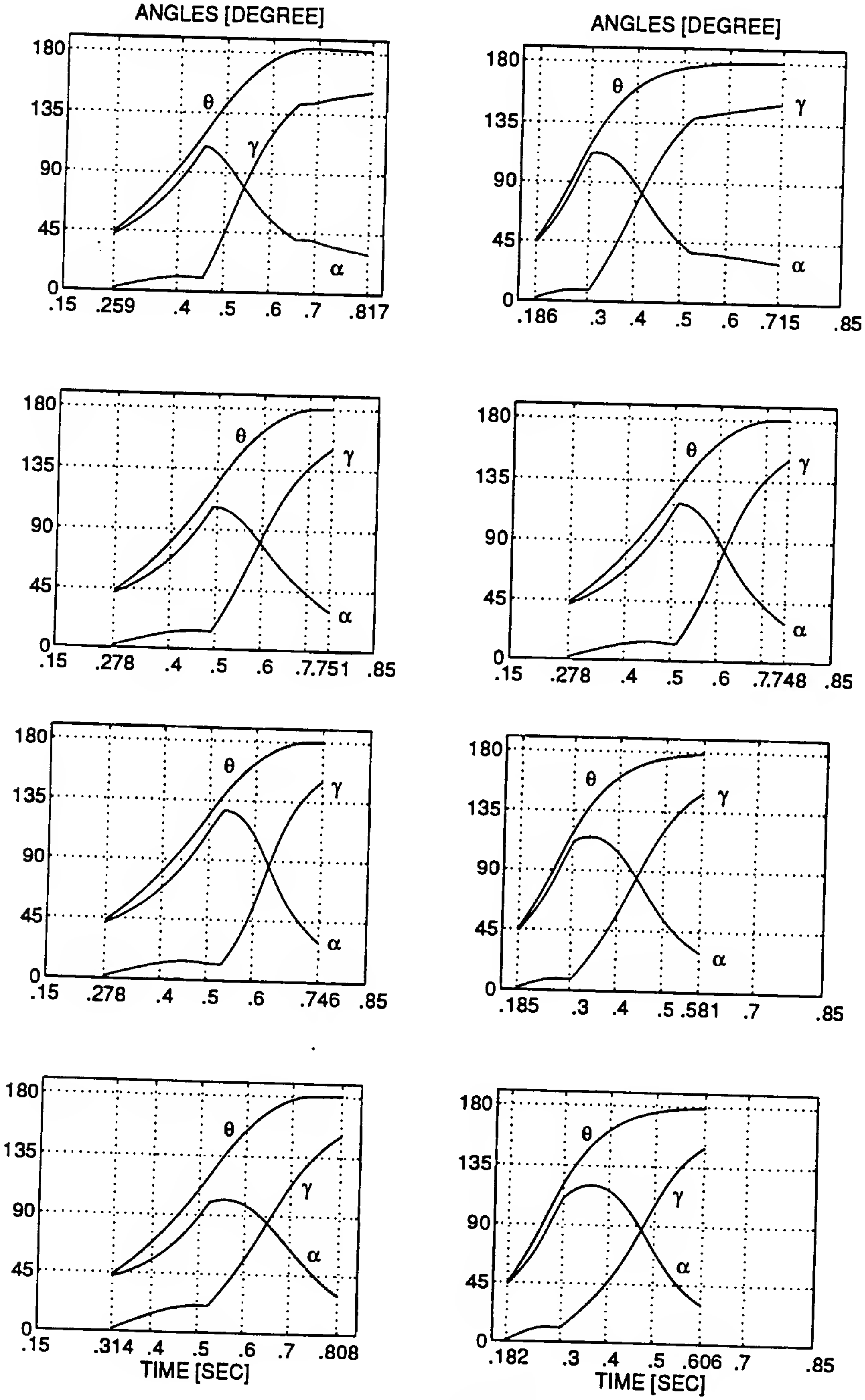

ANGLES [DEGREE]
180
135
90
45
0
θ
γ
α
.15 .259 .4 .5 .6 .7 .817
ANGLES [DEGREE]
180
135
90
45
0
θ
γ
α
.186 .3 .4 .5 .6 .715 .85
180
135
90
45
0
θ
γ
α
.15 .278 .4 .5 .6 .7.751 .85
180
135
90
45
0
θ
γ
α
.15 .278 .4 .5 .6 .7.748 .85
180
135
90
45
0
θ
γ
α
.15 .278 .4 .5 .6 .746 .85
180
135
90
45
0
θ
γ
α
.185 .3 .4 .5 .581 .7 .85
180
135
90
45
0
θ
γ
α
.15 .314 .4 .5 .6 .7 .808
TIME [SEC]
180
135
90
45
0
θ
γ
α
.182 .3 .4 .5 .606 .7 .85
TIME [SEC]

Figure A-6

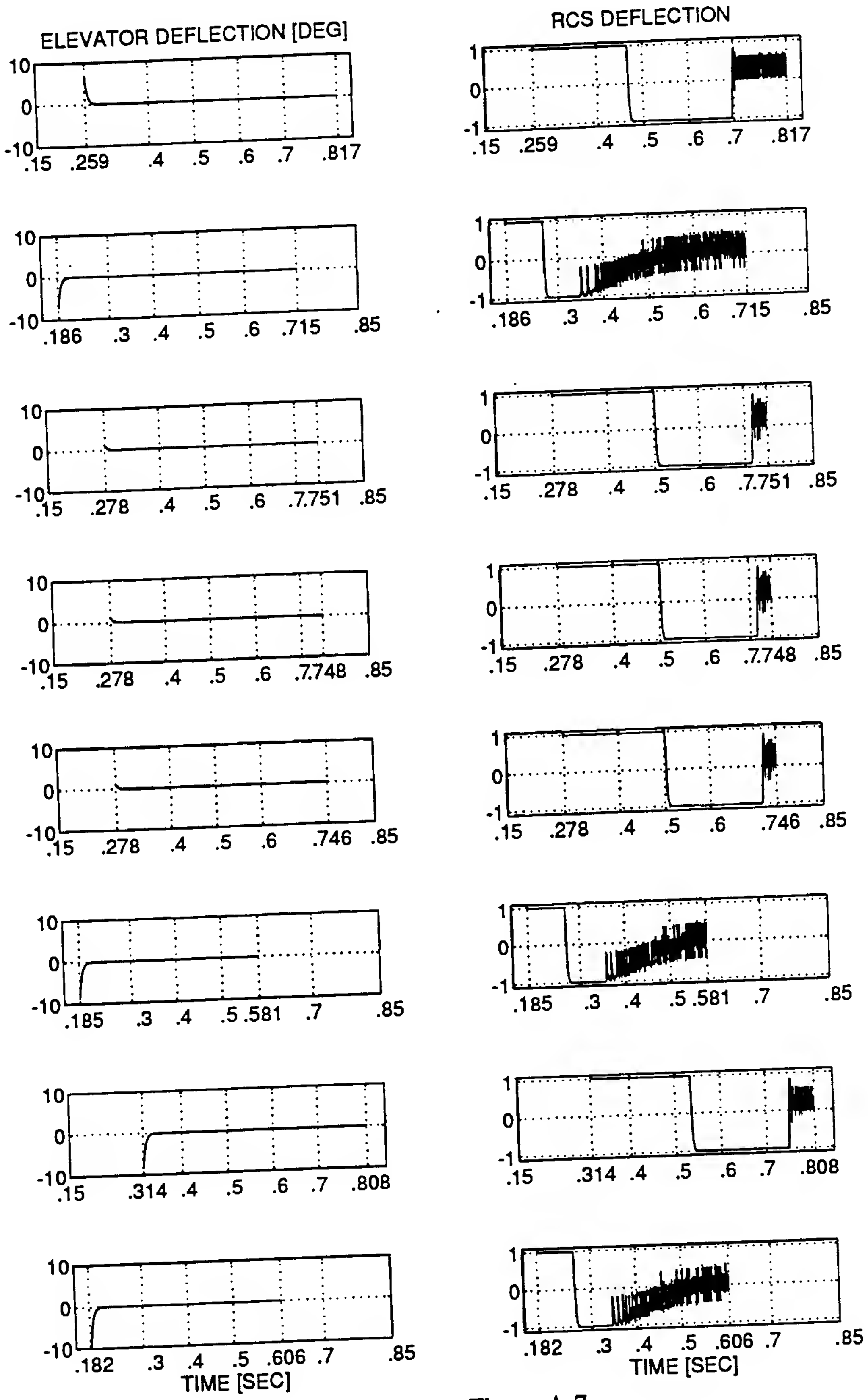

ELEVATOR DEFLECTION [DEG]
RCS DEFLECTION
10
0
-10
.15 .259 .4 .5 .6 .7 .817
1
0
-1
.15 .259 .4 .5 .6 .7 .817
.186 .3 .4 .5 .6 .715 .85
.186 .3 .4 .5 .6 .715 .85
.15 .278 .4 .5 .6 .7.751 .85
.15 .278 .4 .5 .6 .7.751 .85
.15 .278 .4 .5 .6 .7.748 .85
.15 .278 .4 .5 .6 .7.748 .85
.15 .278 .4 .5 .6 .746 .85
.15 .278 .4 .5 .6 .746 .85
.185 .3 .4 .5 .581 .7 .85
.185 .3 .4 .5 .581 .7 .85
.15 .314 .4 .5 .6 .7 .808
.15 .314 .4 .5 .6 .7 .808
.182 .3 .4 .5 .606 .7 .85
.182 .3 .4 .5 .606 .7 .85
TIME [SEC]
TIME [SEC]

Figure A-7

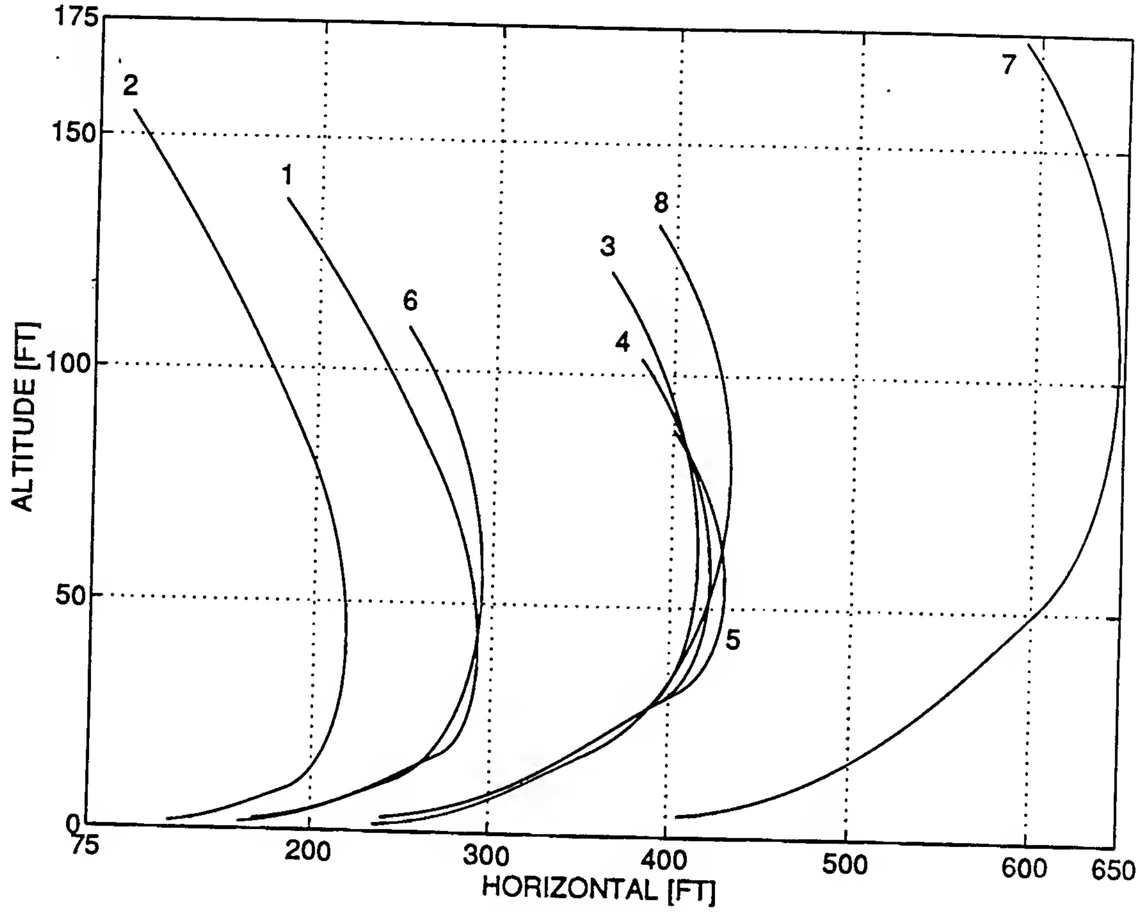

175
150
100
50
0
ALTITUDE [FT]
75
200
300
400
500
600
650
HORIZONTAL [FT]
1
2
3
4
5
6
7
8

Figure A-8

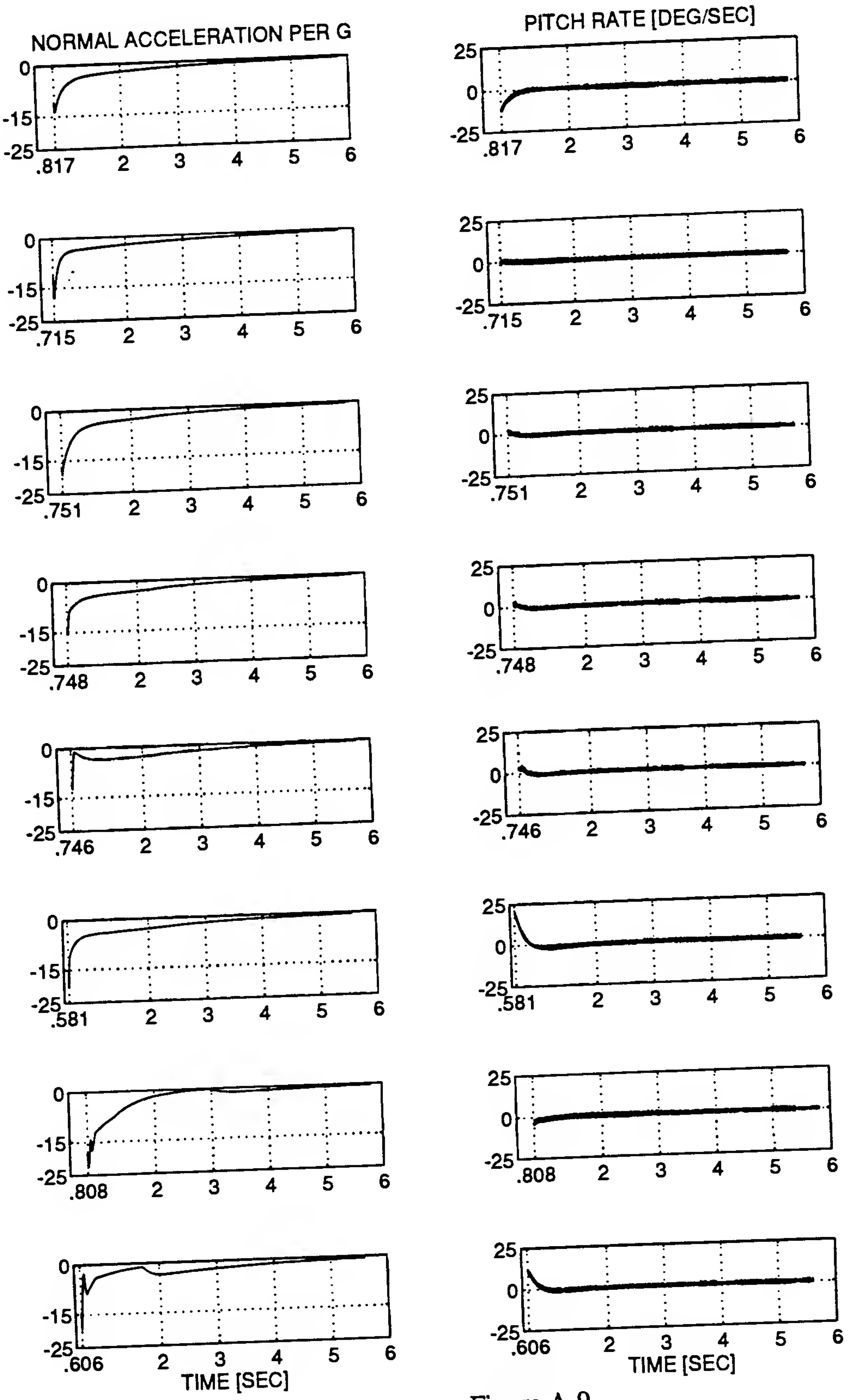

NORMAL ACCELERATION PER G
PITCH RATE [DEG/SEC]
0
-15
-25
25
0
-25
.817
.715
.751
.748
.746
.581
.808
.606
2
3
4
5
6
TIME [SEC]
TIME [SEC]

Figure A-9

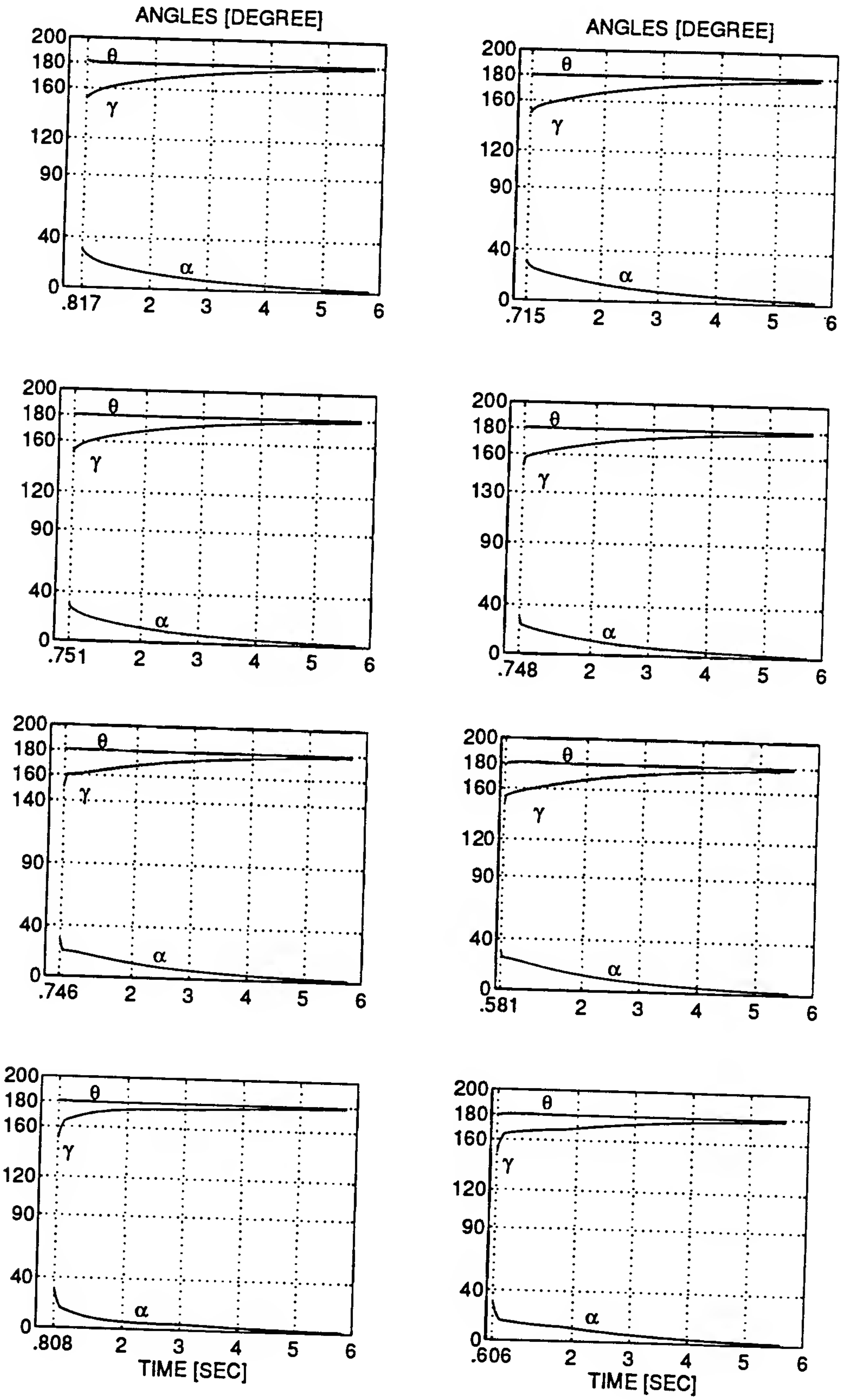
ANGLES [DEGREE]
200
180
160
120
90
40
0
θ
γ
α
.817
2
3
4
5
6
ANGLES [DEGREE]
200
180
160
120
90
40
0
θ
γ
α
.715
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.751
2
3
4
5
6
200
180
160
130
90
40
0
θ
γ
α
.748
2
3
4
5
6
200
180
160
140
90
40
0
θ
γ
α
.746
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.581
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.808
2
3
4
5
6
TIME [SEC]
200
180
160
120
90
40
0
θ
γ
α
.606
2
3
4
5
6
TIME [SEC]

Figure A-10

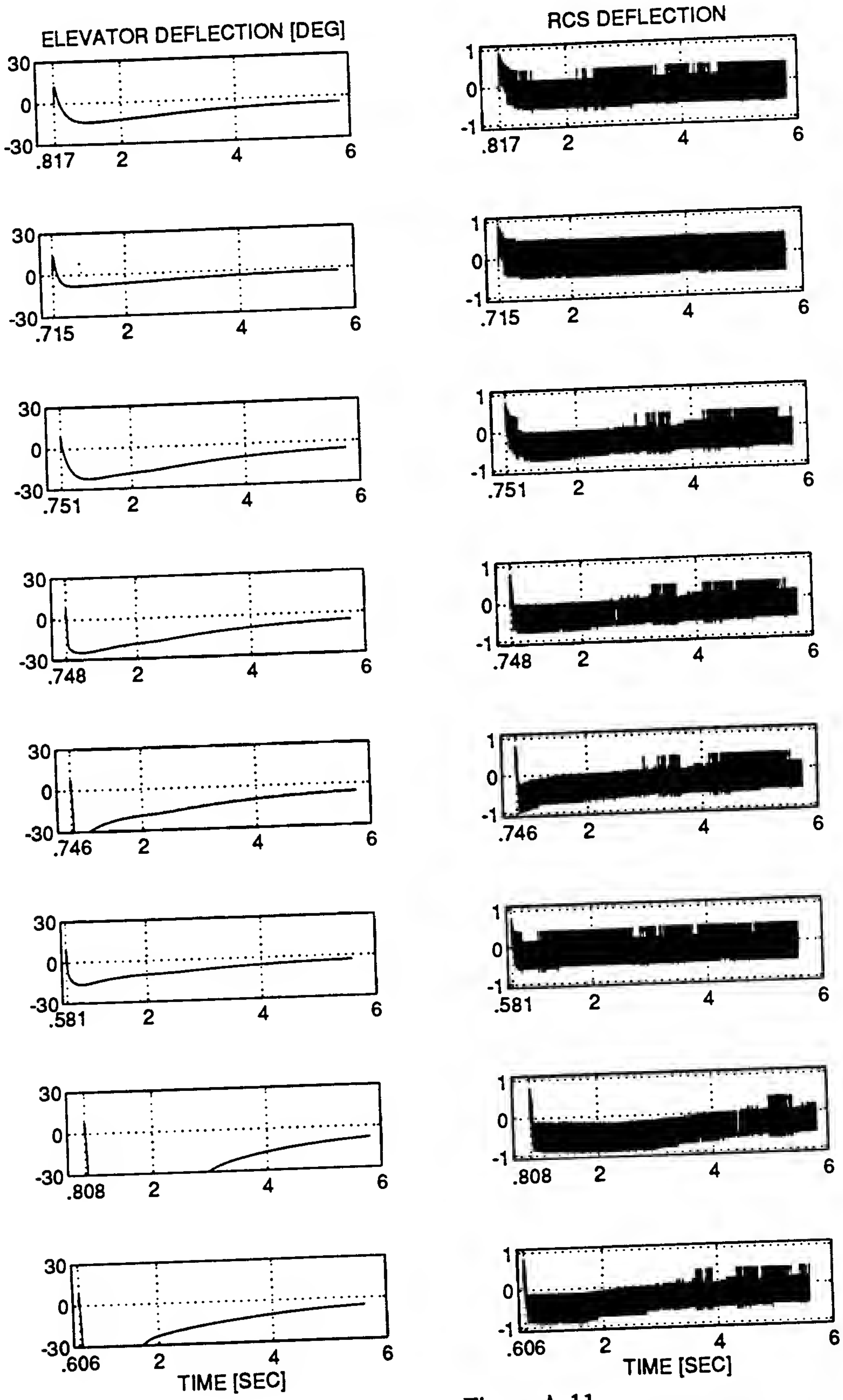

ELEVATOR DEFLECTION [DEG]
RCS DEFLECTION
30
0
-30
1
-1
.817
.715
.751
.748
.746
.581
.808
.606
2
4
6
TIME [SEC]
TIME [SEC]

Figure A-11

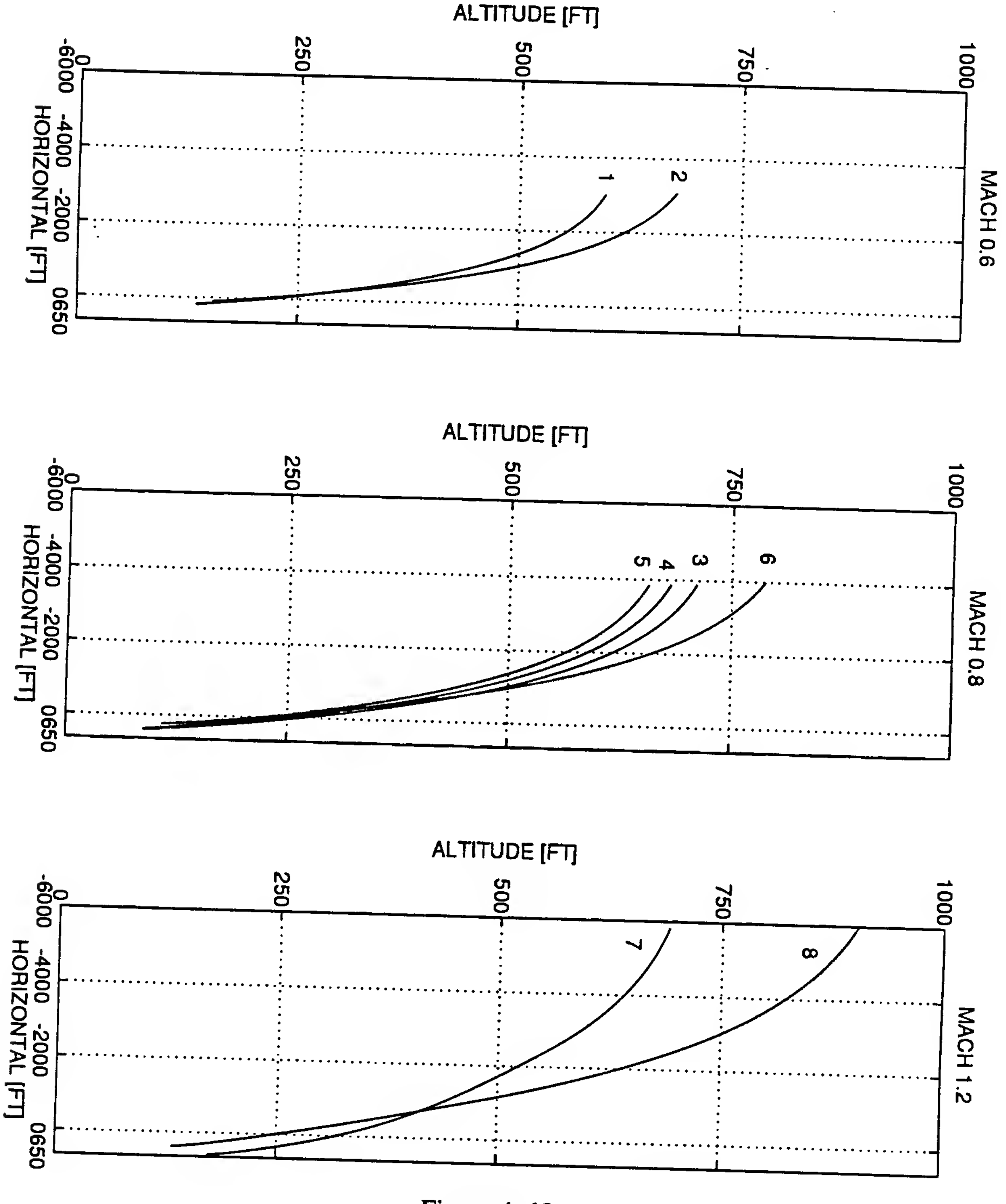

ALTITUDE [FT]
0
250
500
750
1000
-6000
-4000
-2000
0650
HORIZONTAL [FT]
MACH 0.6
1
2
ALTITUDE [FT]
0
250
500
750
1000
-6000
-4000
-2000
0650
HORIZONTAL [FT]
MACH 0.8
3
4
5
6
ALTITUDE [FT]
0
250
500
750
1000
-6000
-4000
-2000
0650
HORIZONTAL [FT]
MACH 1.2
7
8

Figure A-12

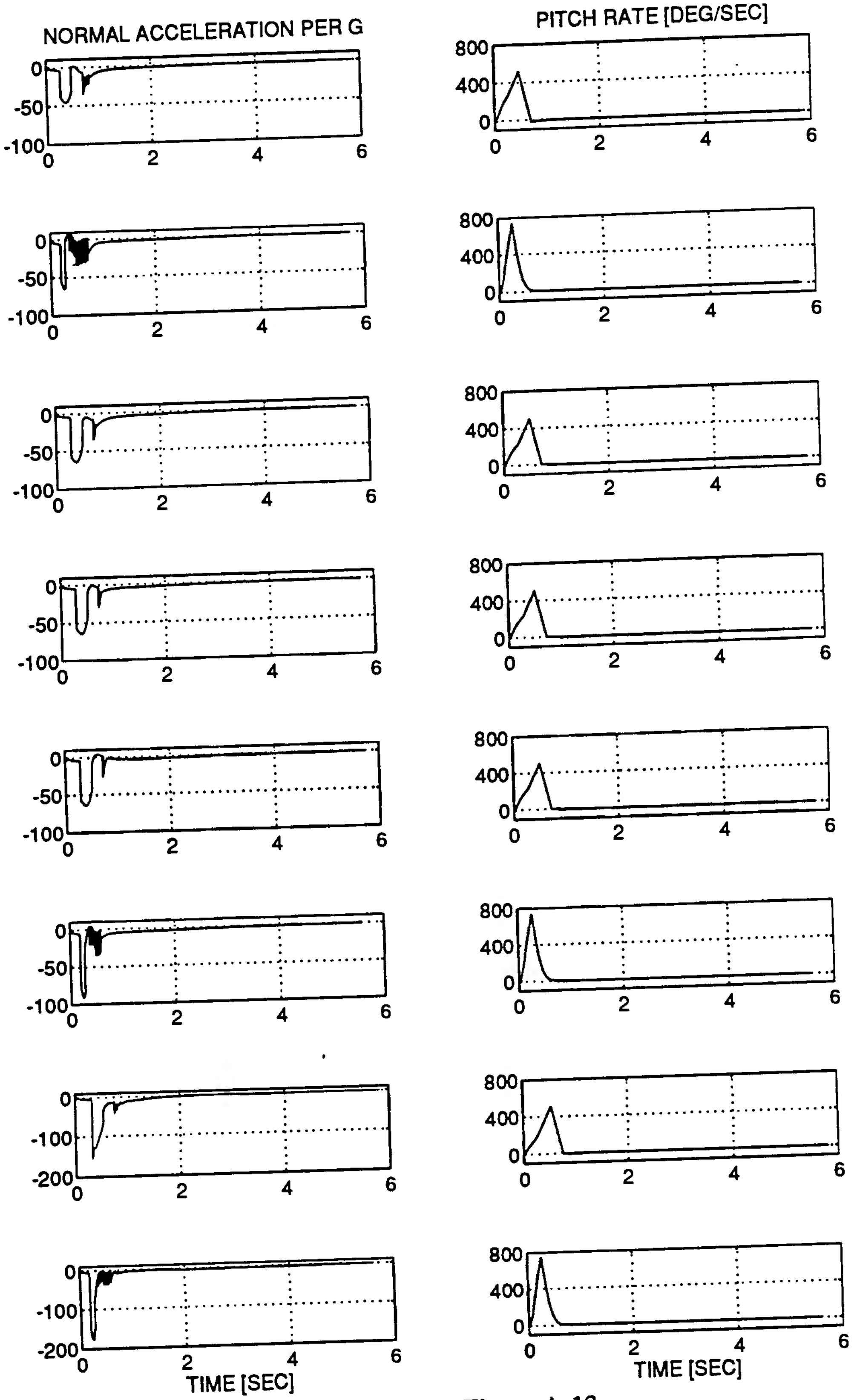

NORMAL ACCELERATION PER G
PITCH RATE [DEG/SEC]
0
-50
-100
-200
800
400
0
2
4
6
TIME [SEC]
TIME [SEC]

Figure A-13

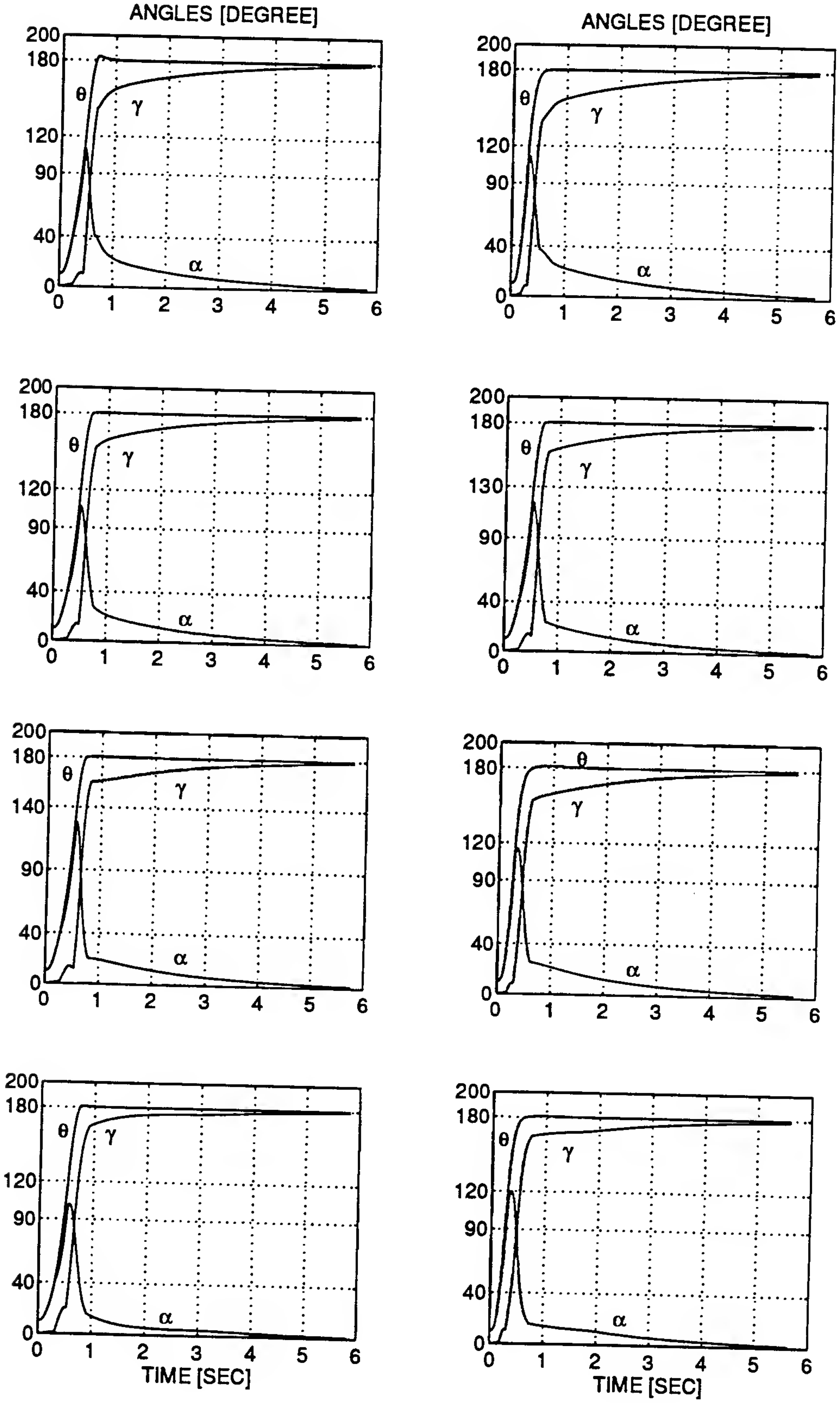

ANGLES [DEGREE]
ANGLES [DEGREE]
200
180
120
90
40
0
0 1 2 3 4 5 6
200
180
120
90
40
0
0 1 2 3 4 5 6
θ
γ
α
200
180
120
90
40
0
0 1 2 3 4 5 6
200
180
130
90
40
0
0 1 2 3 4 5 6
θ
γ
α
200
180
140
90
40
0
0 1 2 3 4 5 6
200
180
120
90
40
0
0 1 2 3 4 5 6
θ
γ
α
200
180
120
90
40
0
0 1 2 3 4 5 6
TIME [SEC]
200
180
120
90
40
0
0 1 2 3 4 5 6
TIME [SEC]

Figure A-14

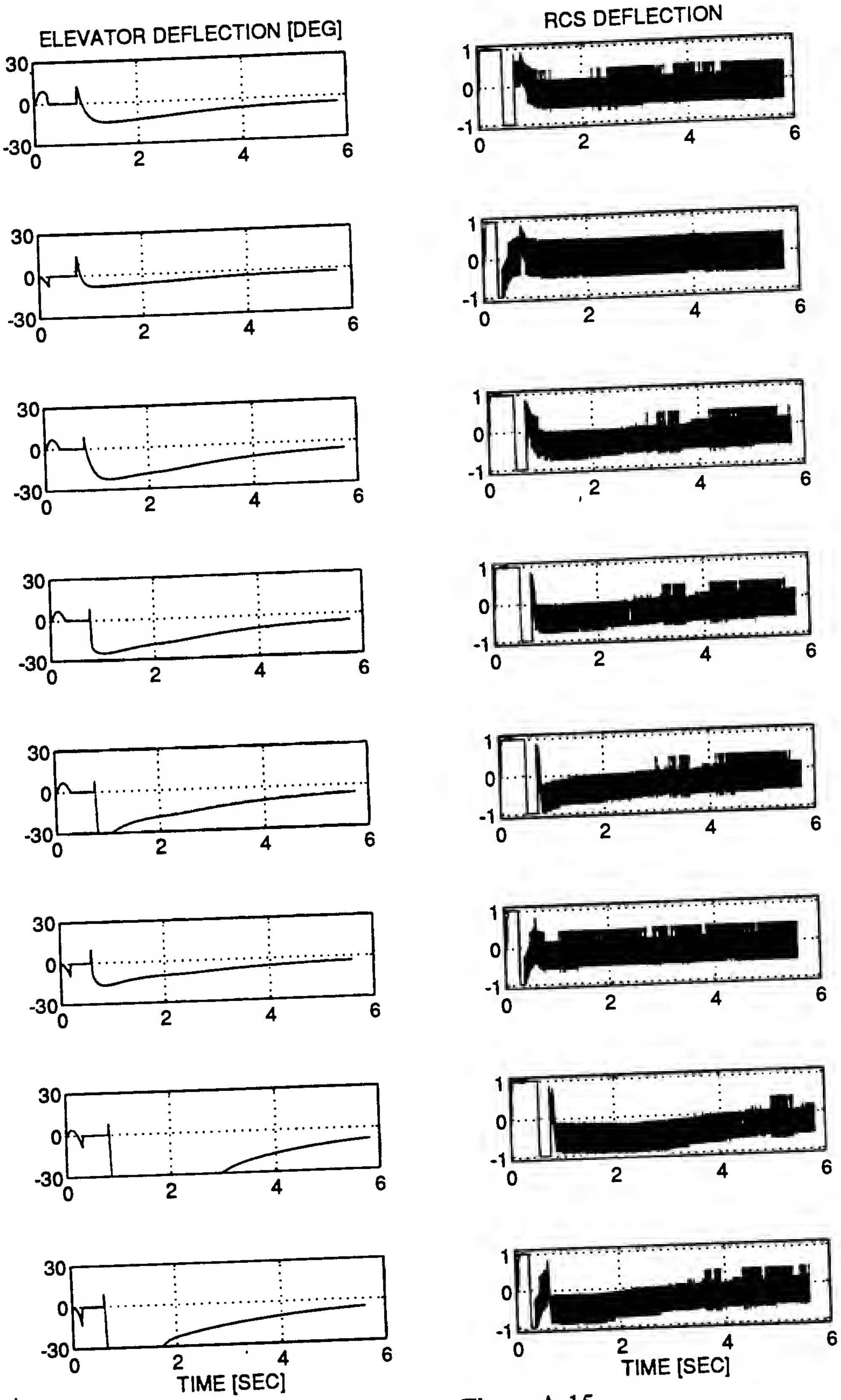

ELEVATOR DEFLECTION [DEG]
RCS DEFLECTION
30
0
-30
1
0
-1
0
2
4
6
TIME [SEC]
TIME [SEC]

Figure A-15

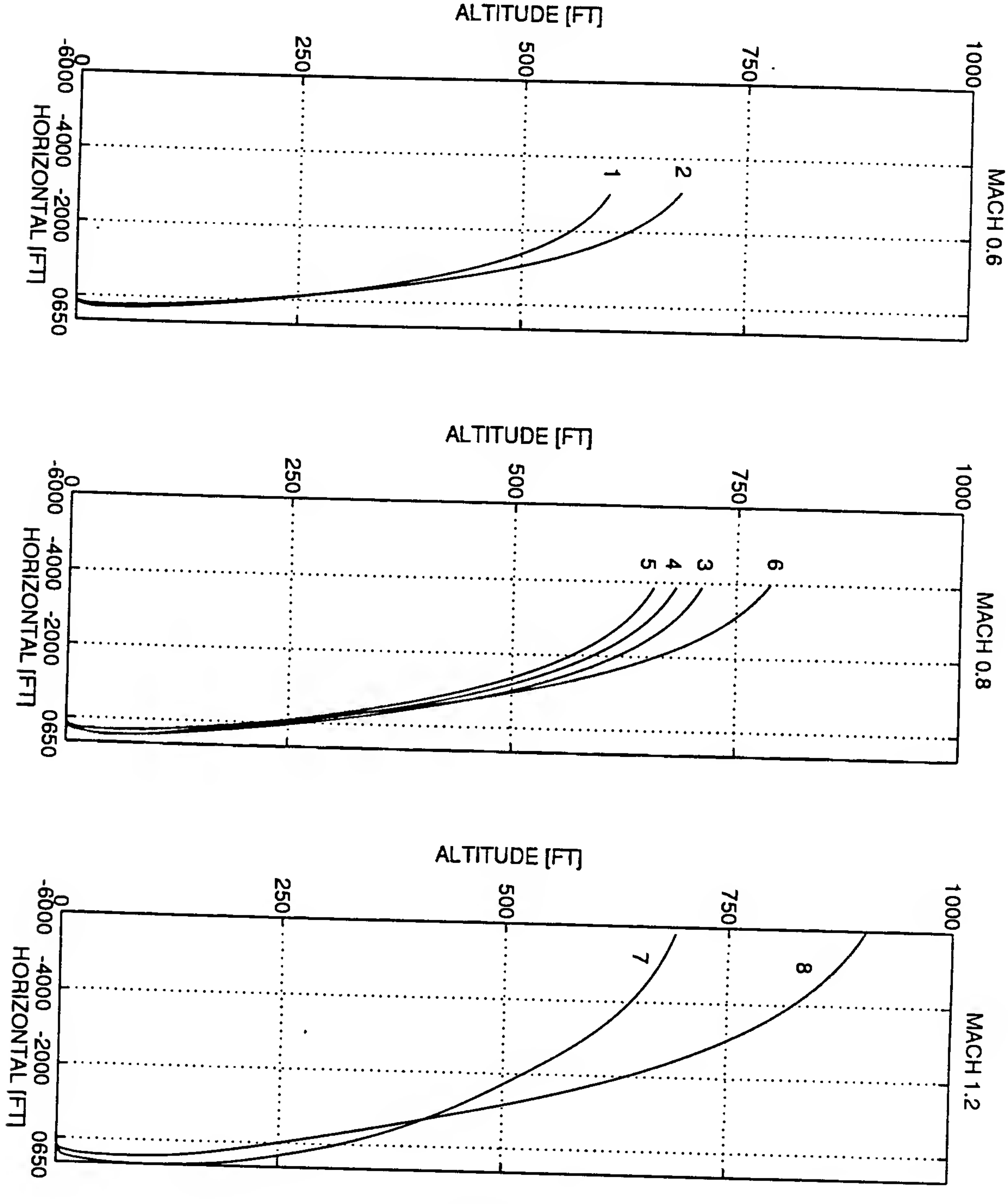

MACH 0.6
ALTITUDE [FT]
0
250
500
750
1000
HORIZONTAL [FT]
-6000
-4000
-2000
0650
1
2
MACH 0.8
ALTITUDE [FT]
0
250
500
750
1000
HORIZONTAL [FT]
-6000
-4000
-2000
0650
3
4
5
6
MACH 1.2
ALTITUDE [FT]
0
250
500
750
1000
HORIZONTAL [FT]
-6000
-4000
-2000
0650
7
8

Figure A-16

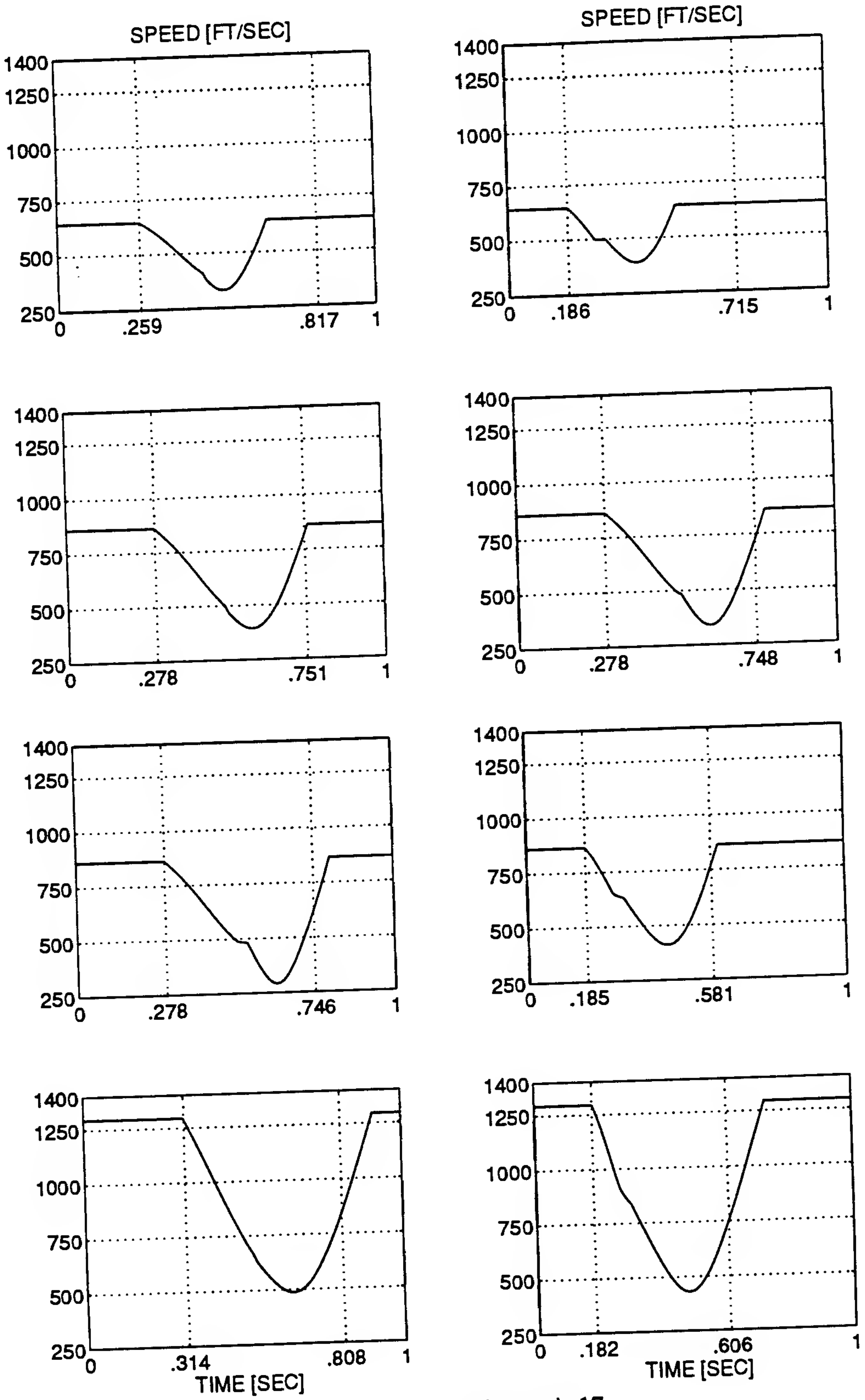

SPEED [FT/SEC]
SPEED [FT/SEC]
1400
1250
1000
750
500
250
0
.259
.817
1
.186
.715
.278
.751
.748
.746
.185
.581
.314
.808
.182
.606
TIME [SEC]
TIME [SEC]

Figure A-17

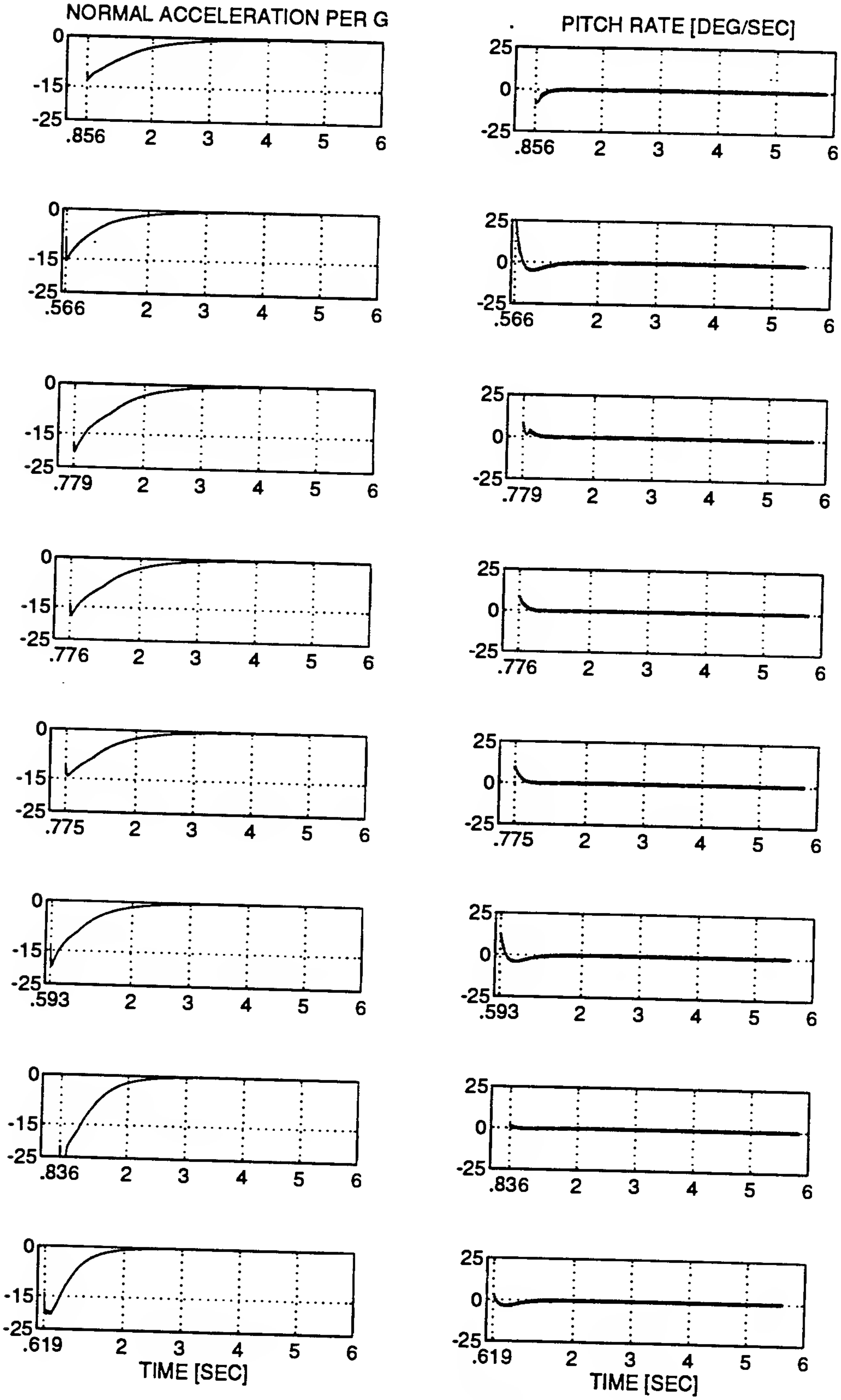

NORMAL ACCELERATION PER G
PITCH RATE [DEG/SEC]
0
-15
-25
25
0
-25
.856
.566
.779
.776
.775
.593
.836
.619
2
3
4
5
6
TIME [SEC]
TIME [SEC]

Figure A-18

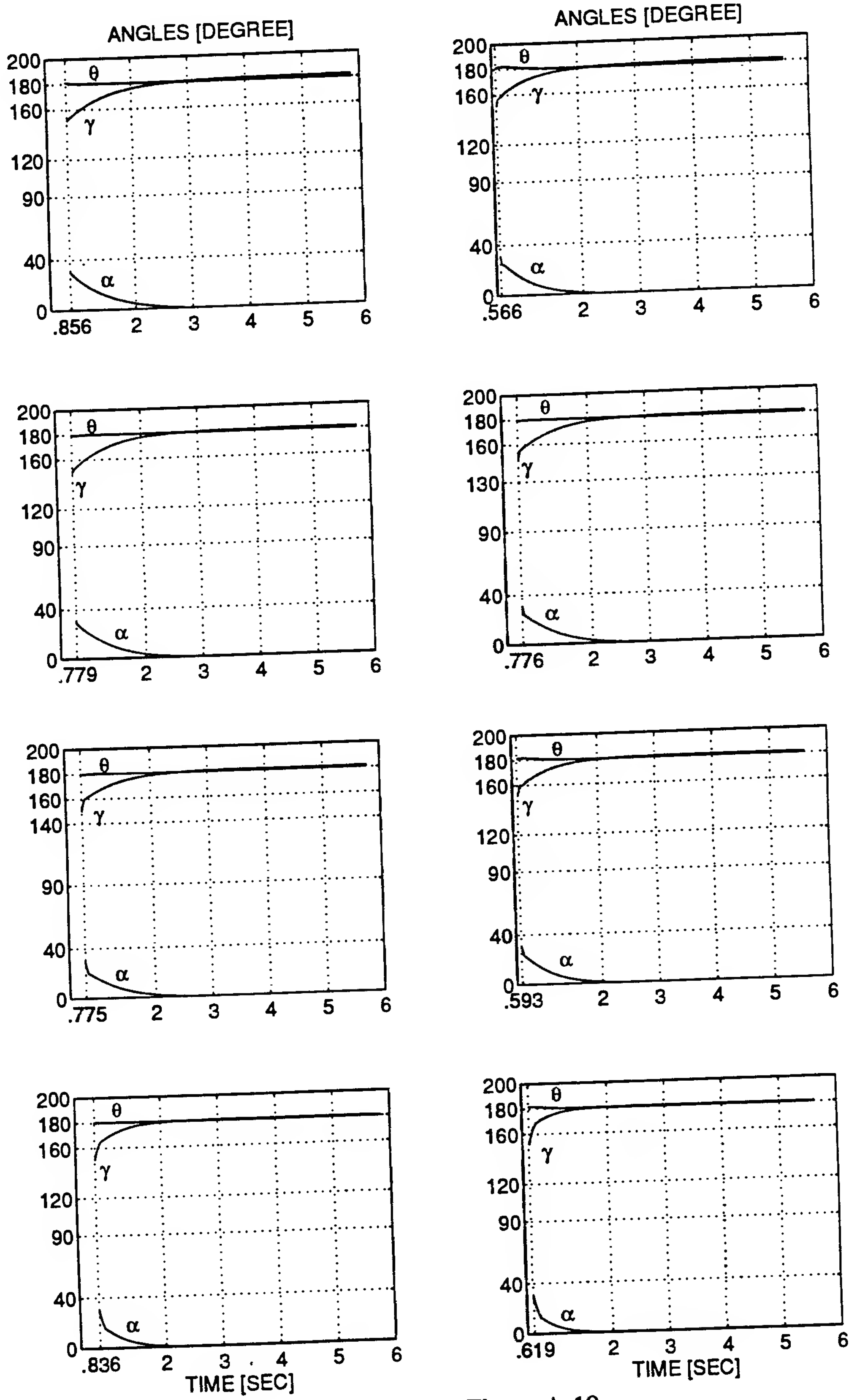

ANGLES [DEGREE]
ANGLES [DEGREE]
200
180
160
120
90
40
0
θ
γ
α
.856
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.566
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.779
2
3
4
5
6
200
180
160
130
90
40
0
θ
γ
α
.776
2
3
4
5
6
200
180
160
140
90
40
0
θ
γ
α
.775
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.593
2
3
4
5
6
200
180
160
120
90
40
0
θ
γ
α
.836
2
3
4
5
6
TIME [SEC]
200
180
160
120
90
40
0
θ
γ
α
.619
2
3
4
5
6
TIME [SEC]

Figure A-19

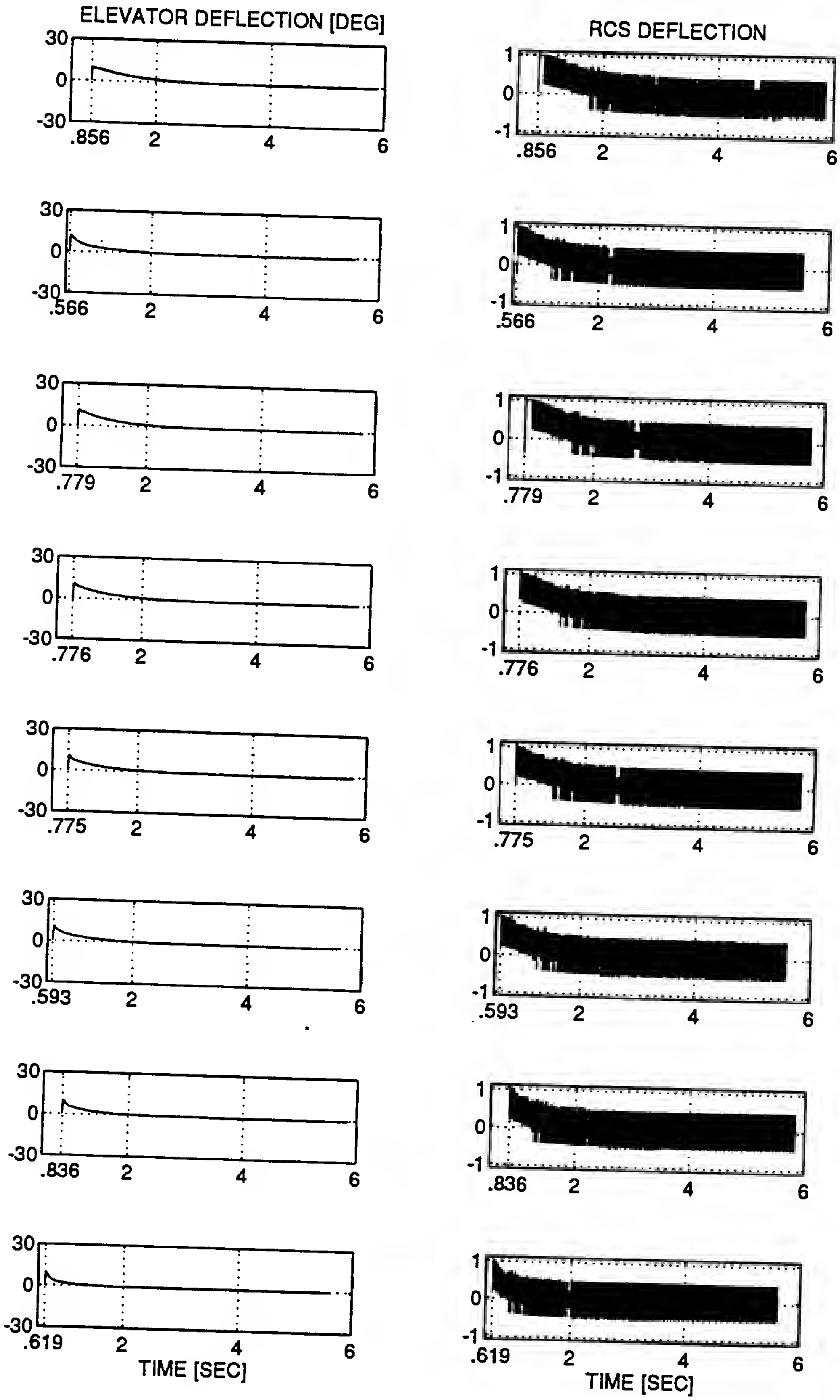

ELEVATOR DEFLECTION [DEG]
RCS DEFLECTION
30
0
-30
1
0
-1
.856
.566
.779
.776
.775
.593
.836
.619
2
4
6
TIME [SEC]
TIME [SEC]

Figure A-20

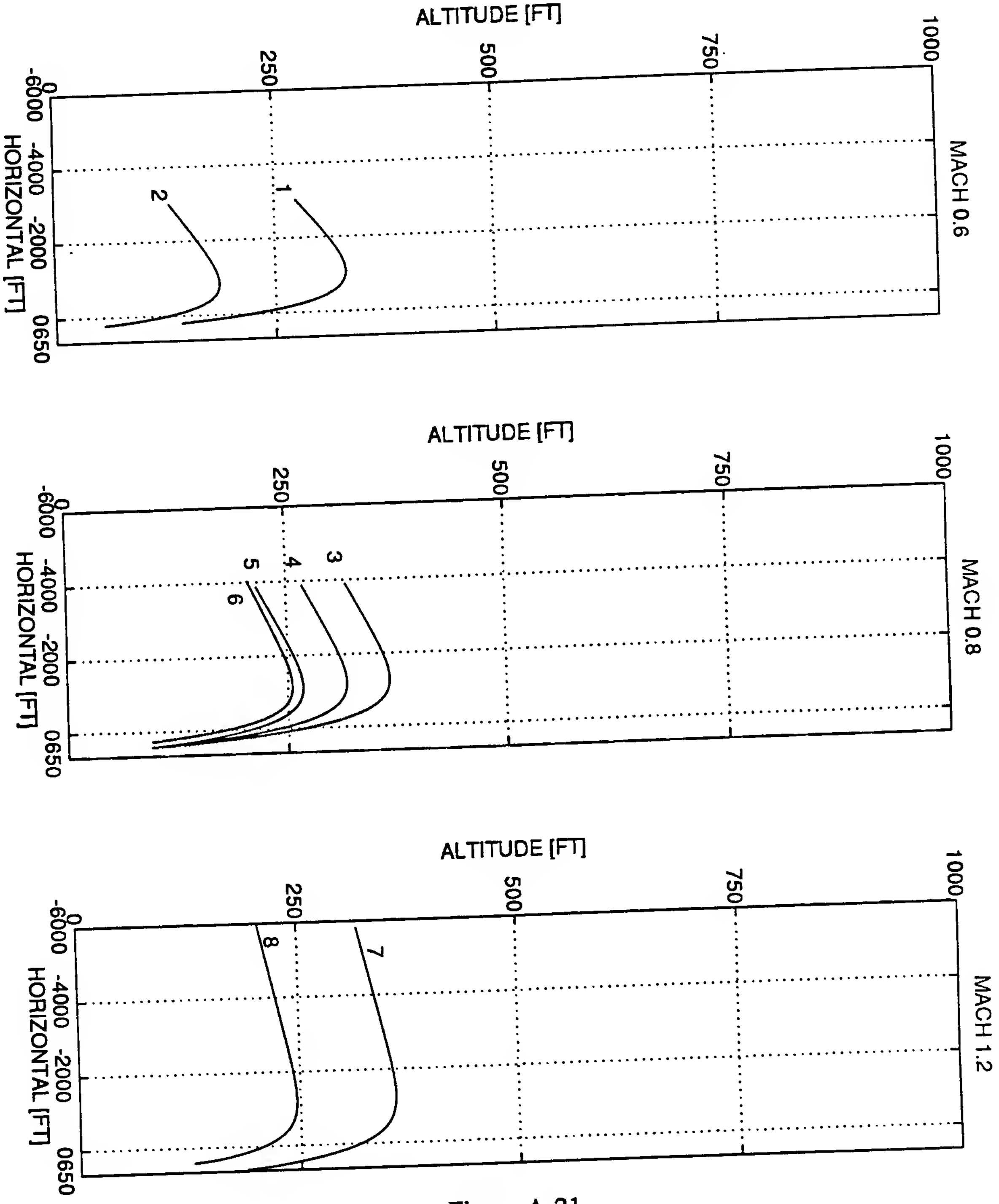

ALTITUDE [FT]
0
250
500
750
1000
MACH 0.6
1
2
-6000
-4000
-2000
0650
HORIZONTAL [FT]
ALTITUDE [FT]
0
250
500
750
1000
MACH 0.8
3
4
5
6
-6000
-4000
-2000
0650
HORIZONTAL [FT]
ALTITUDE [FT]
0
250
500
750
1000
MACH 1.2
7
8
-6000
-4000
-2000
0650
HORIZONTAL [FT]

Figure A-21

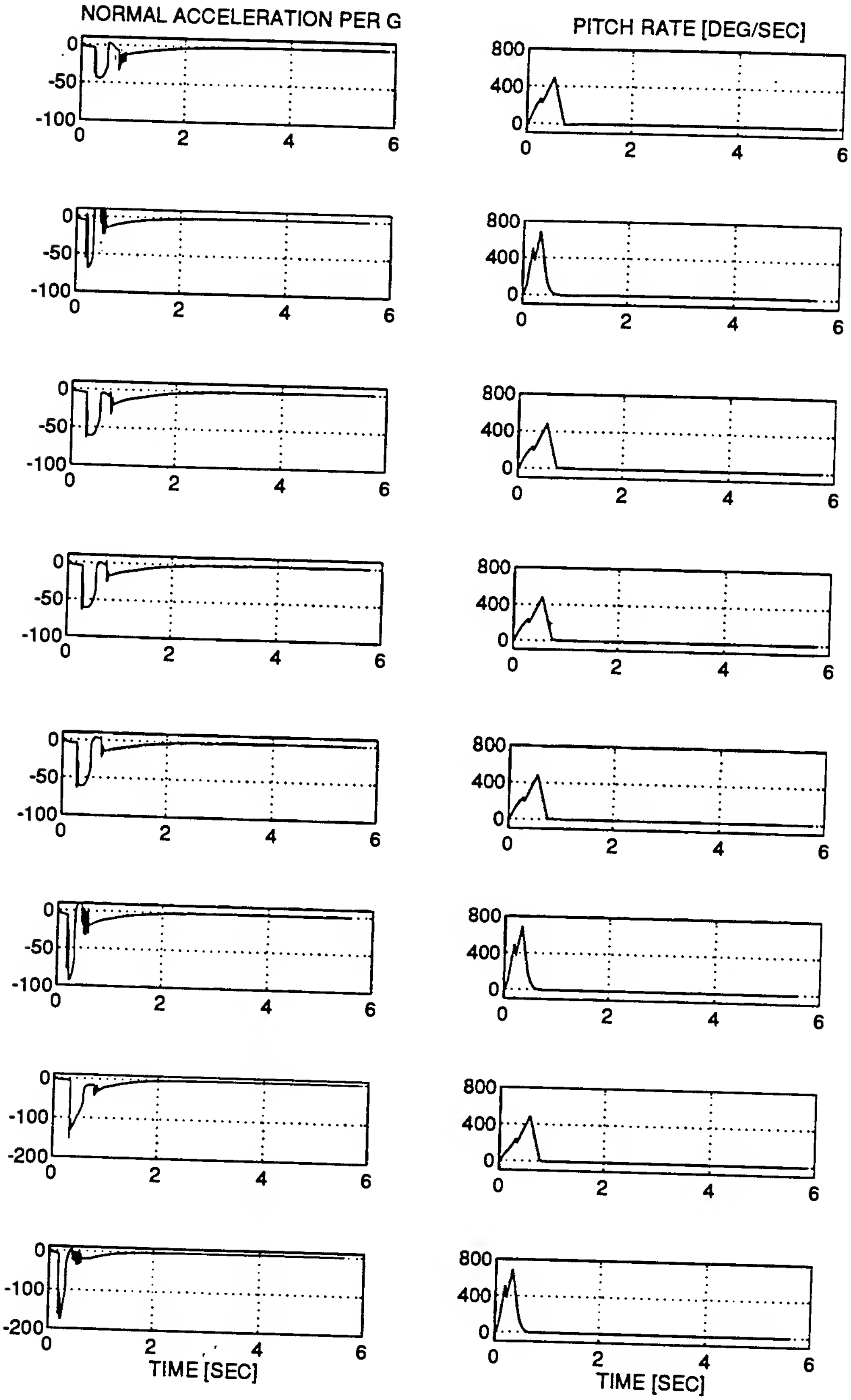

NORMAL ACCELERATION PER G
PITCH RATE [DEG/SEC]
0
-50
-100
-200
800
400
0
2
4
6
TIME [SEC]
TIME [SEC]

Figure A-22

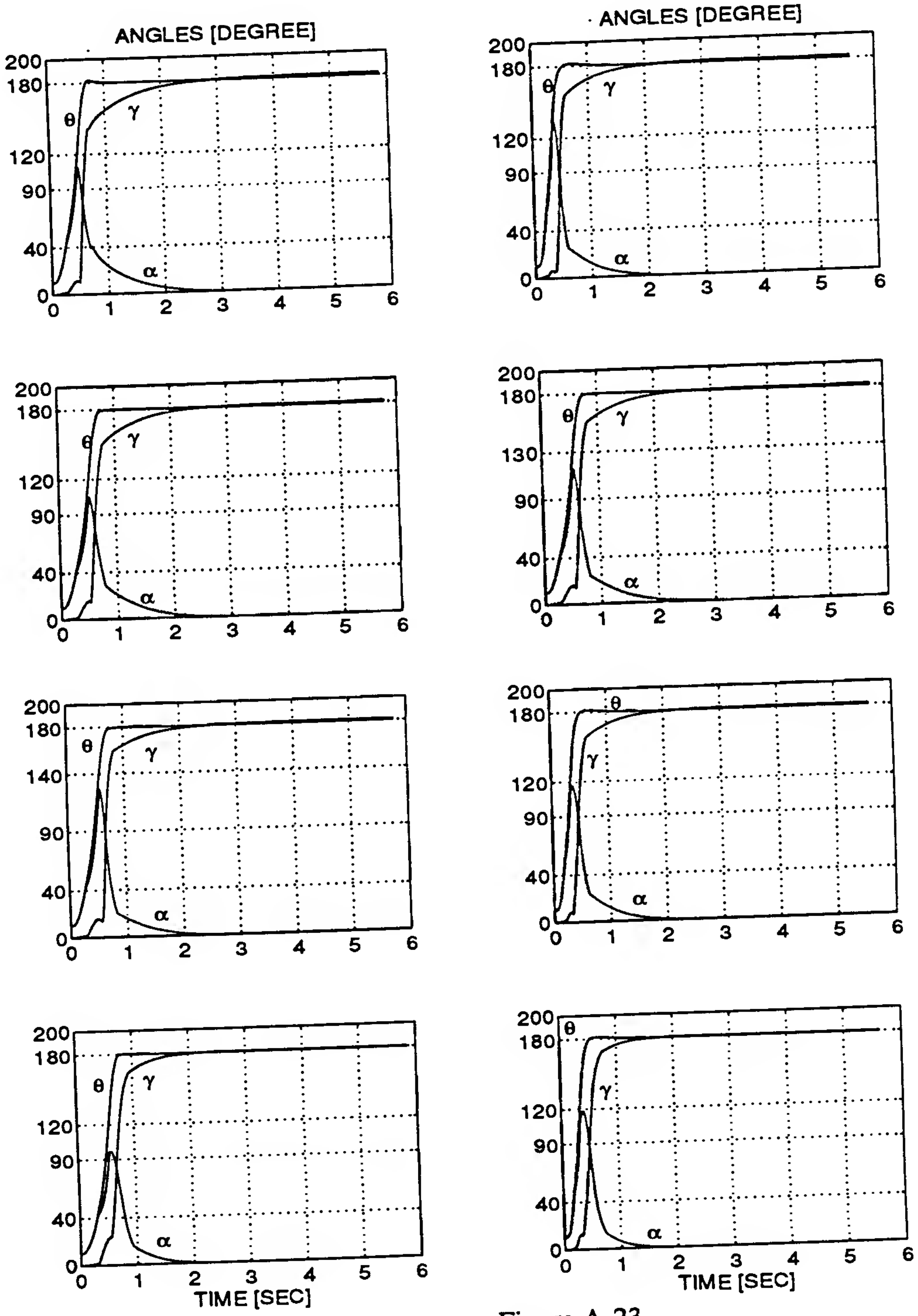

ANGLES [DEGREE]
ANGLES [DEGREE]
200
180
140
130
120
90
40
0
θ
γ
α
0
1
2
3
4
5
6
TIME [SEC]
TIME [SEC]

Figure A-23

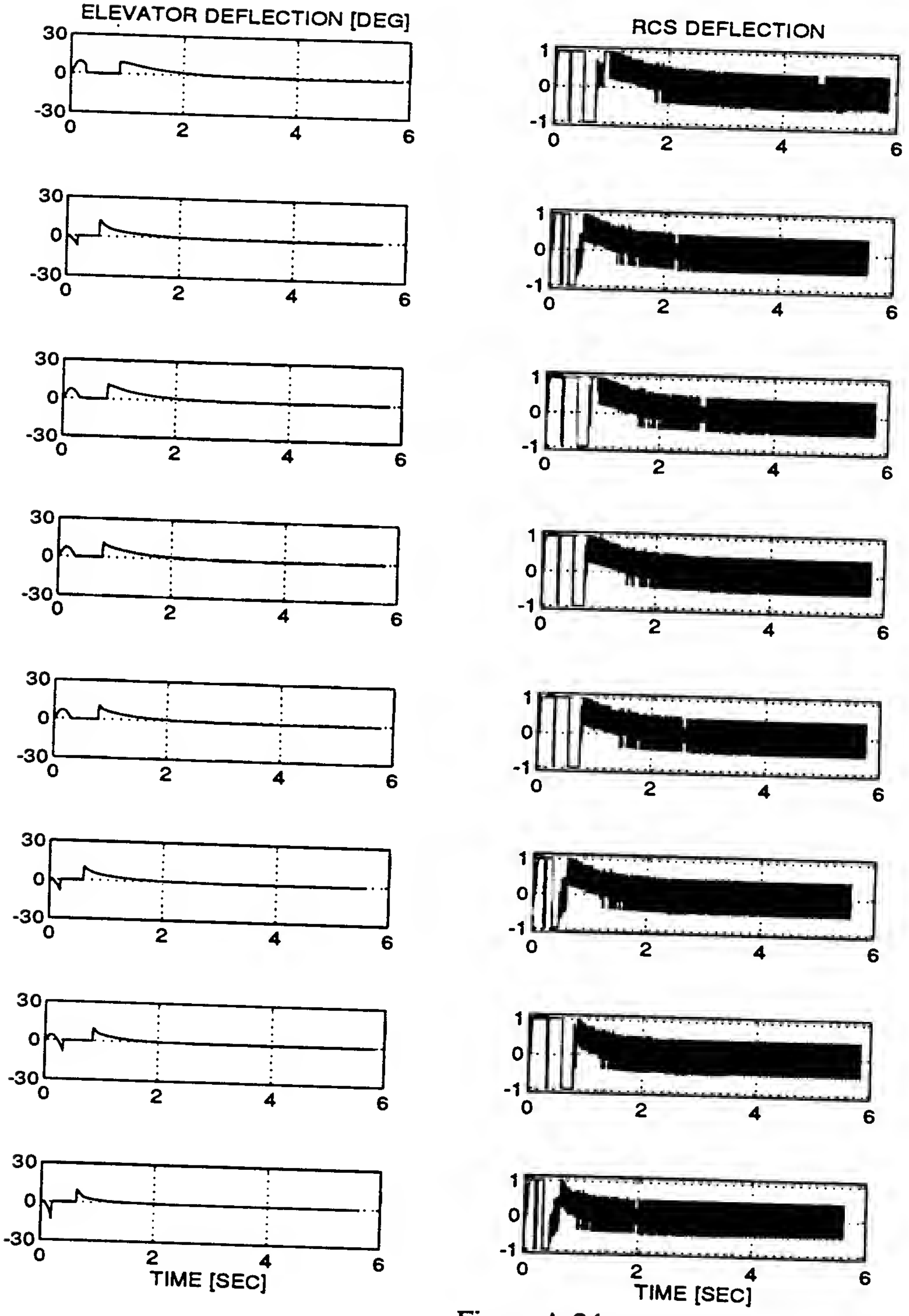

ELEVATOR DEFLECTION [DEG]
RCS DEFLECTION
30
0
-30
1
0
-1
0
2
4
6
TIME [SEC]
TIME [SEC]

Figure A-24

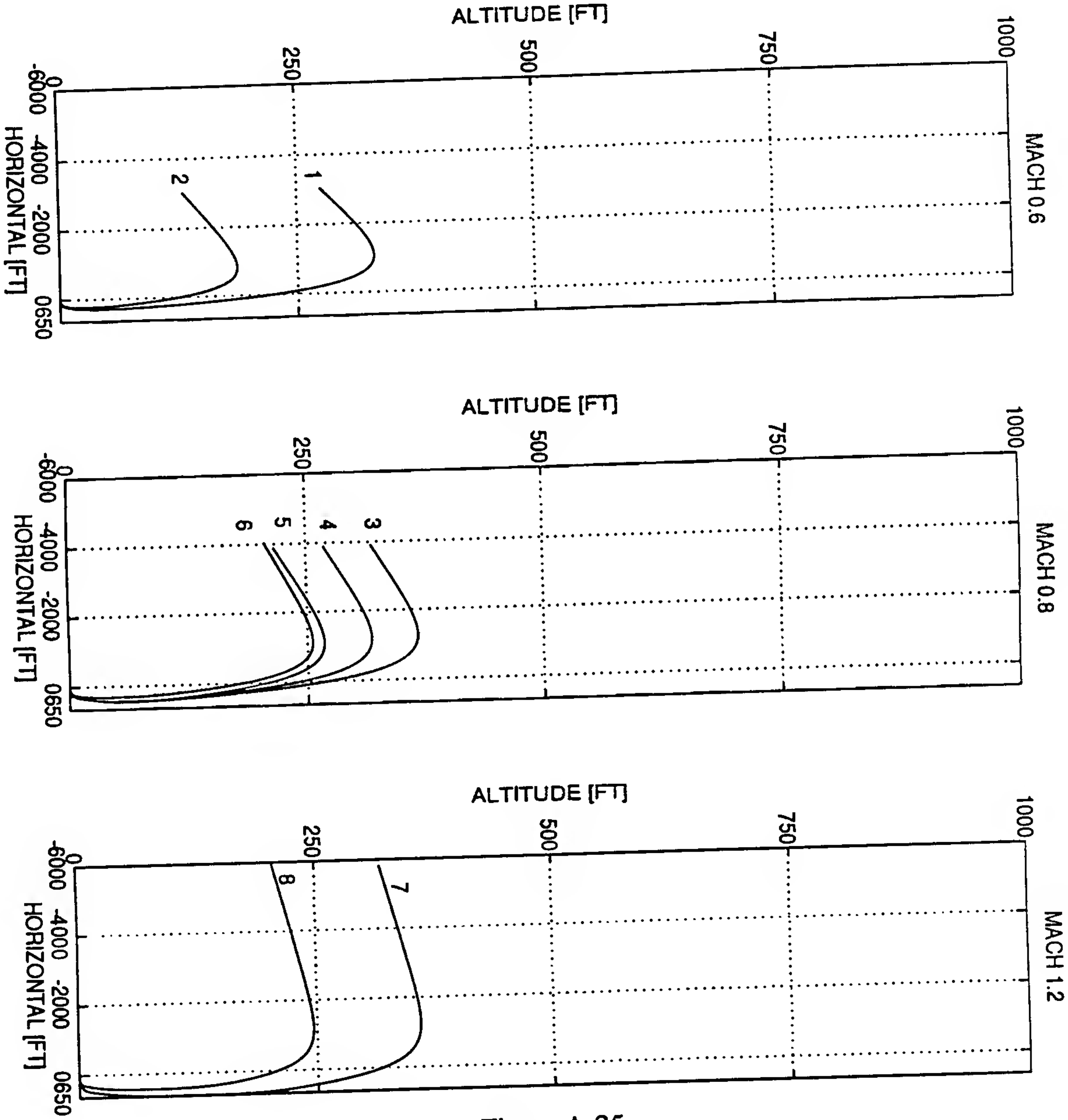

MACH 0.6
ALTITUDE [FT]
0
250
500
750
1000
-6000
-4000
-2000
0650
HORIZONTAL [FT]
1
2
MACH 0.8
ALTITUDE [FT]
0
250
500
750
1000
-6000
-4000
-2000
0650
HORIZONTAL [FT]
3
4
5
6
MACH 1.2
ALTITUDE [FT]
0
250
500
750
1000
-6000
-4000
-2000
0650
HORIZONTAL [FT]
7
8

Figure A-25

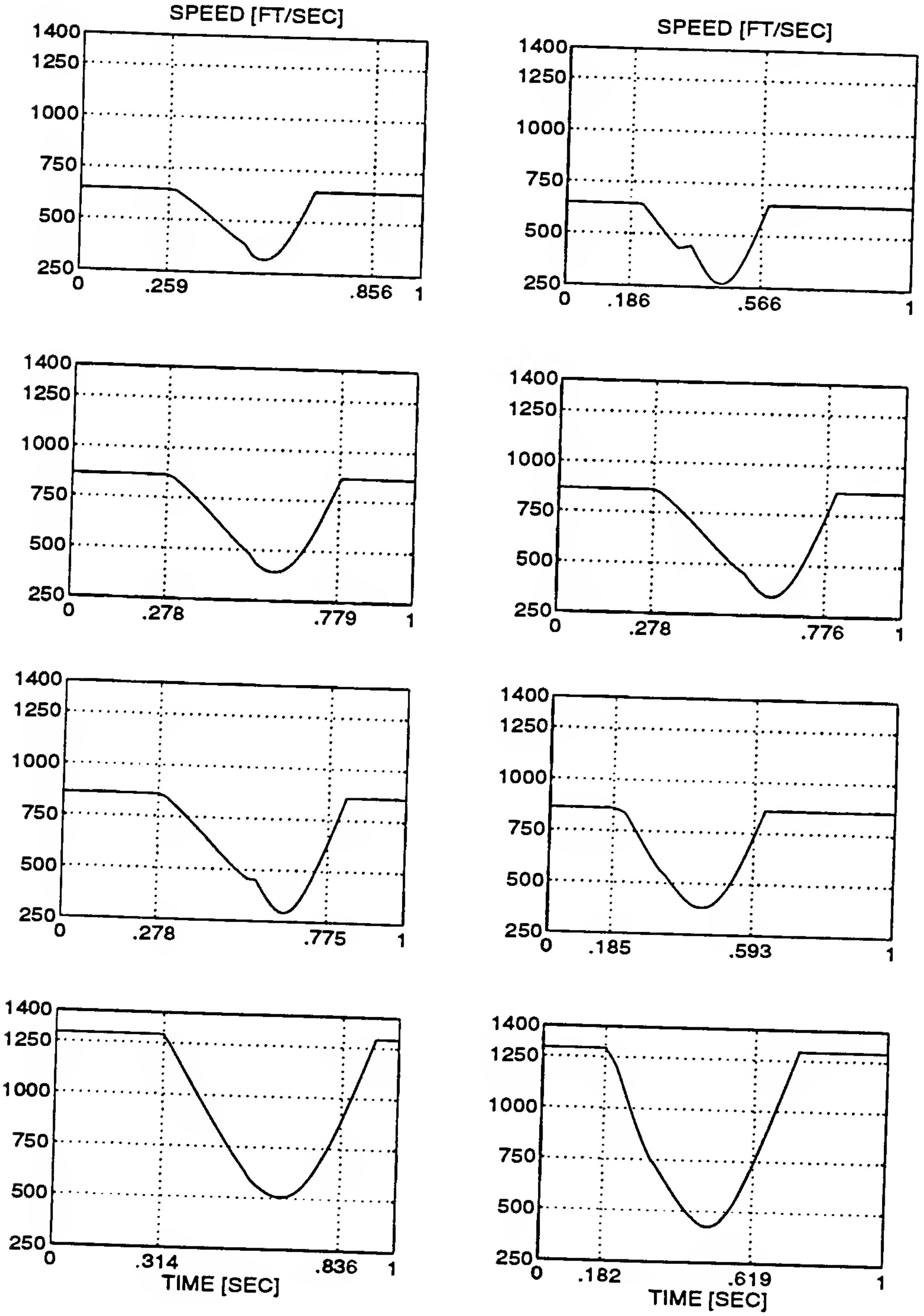

SPEED [FT/SEC]
1400
1250
1000
750
500
250
0
.259
.856
1
SPEED [FT/SEC]
1400
1250
1000
750
500
250
0
.186
.566
1
1400
1250
1000
750
500
250
0
.278
.779
1
1400
1250
1000
750
500
250
0
.278
.776
1
1400
1250
1000
750
500
250
0
.278
.775
1
1400
1250
1000
750
500
250
0
.185
.593
1
1400
1250
1000
750
500
250
0
.314
.836
1
TIME [SEC]
1400
1250
1000
750
500
250
0
.182
.619
1
TIME [SEC]

Figure A-26

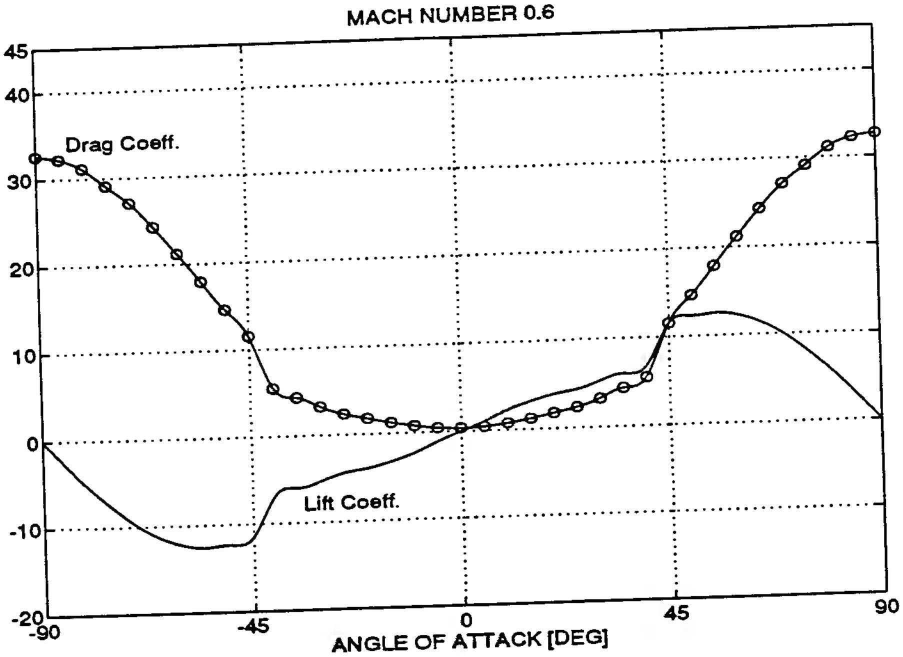


Figure A-27 $C_L$-$C_D$ curves, Mach 0.6

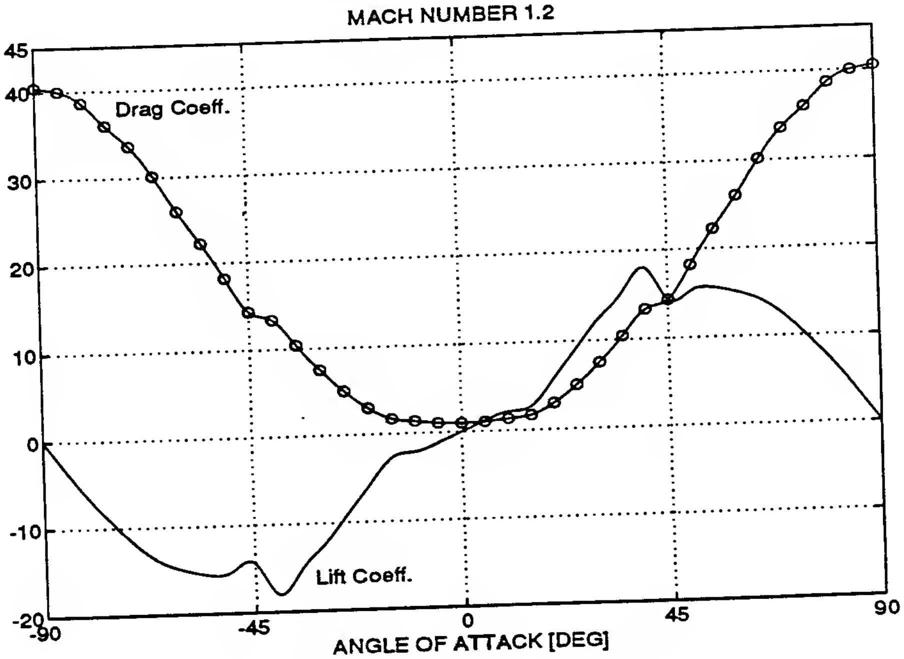


Figure A-28 $C_L$-$C_D$ curves, Mach 1.2

# LASER IMAGING AND RANGING (LIMAR) PROCESSING

Jack S.N. Jean
Assistant Professor
Department of Computer Science and Engineering

Wright State University
Dayton, Ohio 45435

Louis A. Tamburino
Avionics Directorate
Wright Laboratory
Wright-Patterson AFB, Ohio 45433

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Wright State University

December 1993

# LASER IMAGING AND RANGING (LIMAR) PROCESSING

Jack S.N. Jean
Assistant Professor
Department of Computer Science and Engineering
Wright State University

Louis A. Tamburino
Avionics Directorate
Wright Laboratory

## Abstract

The LIMAR (Laser IMaging and Ranging) project is a Wright Laboratory effort to develop an advanced imaging and ranging system for robotics and computer vision applications. LIMAR embodies a concept for the fastest possible three-dimensional camera. It eliminates the conventional scanning processes by producing a registered pair of range and intensity images with data collected from two video cameras. The initial prototype system was assembled and successfully tested at Wright Laboratory's Avionics Directorate in 1992. This prototype LIMAR system used several frame grabbers to capture the demodulated LIMAR image signals from which the range and intensity images were subsequently computed on a general purpose computer. The prototype software did not address the errors which are introduced by differential camera gain, misalignment, and distortion. In last summer, the principal investigator developed algorithms to correct the distortion introduced by using two cameras and designed a special purpose hardware to convert, in real-time, the outputs from the two cameras into a fully registered range and intensity image. During the project period, the detailed design of the processing system has been finished and three hardware boards have been implemented and tested.

LASER IMAGING AND RANGING (LIMAR) PROCESSING

Jack S.N. Jean and Louis A. Tamburino

## 1 Introduction

The LIMAR (Laser IMaging and Ranging) project is a Wright Laboratory effort to develop an advanced imaging and ranging system for robotics and computer vision applications. LIMAR, the invention of Dr. Louis A. Tamburino of Wright Laboratory and Dr. John Taboada of the USAF School of Aerospace Medicine, embodies a concept for the fastest possible three-dimensional camera. It eliminates the conventional scanning processes by producing a registered pair of range and intensity images with data collected from two video cameras. The initial prototype system was assembled and successfully tested at Wright Laboratory's Avionics Directorate in 1992. The prototype LIMAR system used several frame grabbers to capture the demodulated LIMAR image signals from which the range and intensity images were subsequently computed on a general purpose computer.

The intent of this research effort is to design and implement a customized LIMAR processing system to generate both the range and intensity images in real time. This new ability is needed to facilitate future robotic and automatic vision applications. The processor design also takes into consideration error correction for camera distortions and misalignments. Error correction was not explored in the original LIMAR prototype and was investigated in the summer of 1992. The result was the development of an alignment algorithm. This report first describes the computations required for the processing and an algorithm used to reduce computations. It then describes a LIMAR processing system and several hardware boards implemented.

**LIMAR Device Overview** As shown in Figure 1, the LIMAR device contains a laser, a polarization modulator, a beam splitter, and a processing unit. The laser shines light on the object which reflects the light back to the modulator. The modulator changes the polarization of the returned light. The polarization change depends on when the light reaches the modulator, or equivalently, depends on the distance between the object and the LIMAR device. Note that the

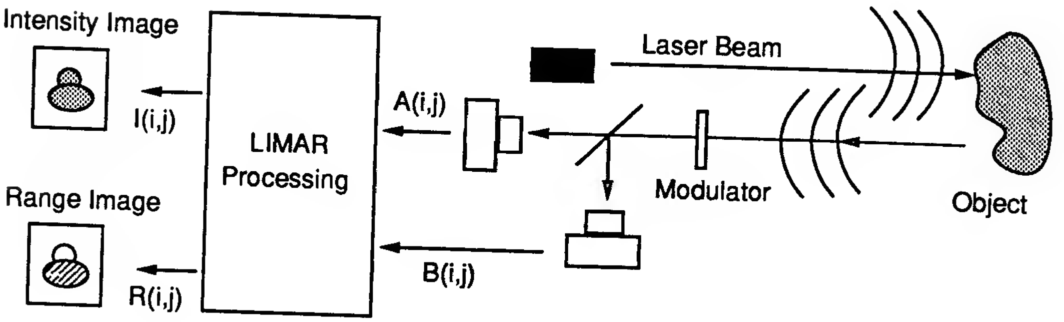


Figure 1: LIMAR system overview

light considered here is not a single spot but an area where the polarization of each single spot can be different from that of others. The beam splitter separates the light into two bundles of light, each captured by a camera, and the degree of separation is a function of the polarization of each spot. With the separation, some computation can be performed to extract the polarization for each single spot, and therefore, to calculate the distance of each spot on the object surface to the LIMAR device. In addition, the intensity, or the reflectivity, of each object spot can be obtained.

In an ideal case, the intensity image $\mathbf{I}(i,j)$ and the range image $\mathbf{R}(i,j)$ are as follows.

$$\begin{aligned} \mathbf{I}(i,j) &= \mathbf{A}(i,j) + \mathbf{B}(i,j) \qquad (1) \\ \mathbf{R}(i,j) &= c\,\mathbf{f}(\, tan^{-1}\sqrt{\frac{\mathbf{A}(i,j)}{\mathbf{B}(i,j)}}\,) + d \qquad (2) \end{aligned}$$

where $\mathbf{A}(i,j)$ and $\mathbf{B}(i,j)$ are the two captured images from cameras, $c$ and $d$ are two constants, and the function $f(\cdot)$ is the inverse of the characteristic function associated with the modulator. In other words, the two output images can be computed pixel by pixel from the two input images in the ideal case.

In the 1992 summer, the distortion introduced by using two cameras was studied. The resulting alignment algorithm and its computational requirements are summarized in Section 2. Also included in the section is a recursive polynomial evaluation technique which is used to remove the multiplications so to simplify implementation. A LIMAR processing system and several hardware boards resulting from the project are described in Section 3.

# 2 Computational Requirements

## 2.1 Image Alignment

Because the two input images are grabbed with two cameras, they may not be aligned well with respect to each other. Furthermore, the two cameras may introduce different distortions. So there is a need to first align each of them to an "ideal" image before the pixel by pixel operations as indicated in equations 1 and 2 can be performed.

The alignment process contains two operations, *position mapping* and *interpolation*. The position mapping transforms a pixel location $(u, v)$ to a new location $(x, y)$ after alignment. It was modeled as a bi-variate polynomials as follows.

$$x = a_0 + a_1 u + a_2 v + a_3 u^2 + a_4 uv + a_5 v^2 + a_6 u^3 + a_7 u^2 v + a_8 uv^2 + a_9 v^3 + \cdots \quad (3)$$

$$y = b_0 + b_1 u + b_2 v + b_3 u^2 + b_4 uv + b_5 v^2 + b_6 u^3 + b_7 u^2 v + b_8 uv^2 + b_9 v^3 + \cdots \quad (4)$$

Since the new location $(x, y)$ may not be a meaningful pixel location, e.g., $x$ or $y$ may not be integer, bilinear interpolation is used. In the summer of 1992, it was found that third-degree polynomials are good enough for alignment purposes and the polynomial coefficients can be computed with a least square estimation algorithm.

## 2.2 Polynomial Evaluation

The evaluation of two third-degree bi-variate polynomials for each pixel is pretty expensive in terms of computation. For images of size M (rows) by N (columns), the polynomials x(u,v) and y(u,v) in equations 3 and 4 are to be evaluated for all the (u,v)'s in {(u,v) | u ∈ {0, 1, 2, ..., N-1} and v ∈ {0, 1, 2, ..., M-1}}. In our case, M=480 and N=640. Since there are 30 frames of images per second, the computation requires 18,432,000 (=2x480x640x30) polynomial evaluations per second. As a result, a fast way to evaluate the polynomials is necessary. The technique, recursive evaluation, takes only three additions per pixel for the evaluation of a third-degree bi-variate polynomial. Numerical stability issues are also taken into consideration in the technique.

To illustrate the approach, the evaluation of a third-degree single-variable polynomial is first presented. The result is then applied to the bi-variate case.

**Single-variable Polynomial Evaluation** Let $y_3(x) = c_0 + c_1x + c_2x^2 + c_3x^3$, where x is in the range, 0, 1, 2, ..., N-1. Instead of performing the N evaluations separately, which takes at least 3N multiplications and 3N additions, a recursive formulation can be used to reduce the number of computations. That is,

$$\begin{aligned} y_3(x+1) &= y_3(x) + y_2(x) \\ y_3(0) &= c_0 \end{aligned}$$

where $y_2(x)$ is a second-degree polynomial which is equal to $(c_1 + c_2 + c_3) + (2c_2 + 3c_3)x + 3c_3x^2$. Since the evaluation of a second-degree polynomial takes at least two multiplications and two additions, the required computations are reduced to 2N multiplications and 3N additions (plus the three multiplications and three additions in obtaining the coefficients of $y_2(x)$). For large N, the 6N computations have been reduced to 5N.

Similarly, recursive formulation can be applied to $y_2(x)$, or

$$\begin{aligned} y_2(x+1) &= y_2(x) + y_1(x) \\ y_2(0) &= c_1 + c_2 + c_3 \end{aligned}$$

and then be applied to $y_1$ $[= (2c_2 + 6c_3) + 6c_3x]$ such that

$$\begin{aligned} y_1(x+1) &= y_1(x) + 6c_3 \\ y_1(0) &= 2c_2 + 6c_3 \end{aligned}$$

In summary, the following recursive formula can be used to evaluate all the $y_3(x)$ values sequentially.

$$\begin{aligned} y_3(x+1) &= y_3(x) + y_2(x) \\ y_2(x+1) &= y_2(x) + y_1(x) \\ y_1(x+1) &= y_1(x) + 6c_3 \end{aligned}$$

given the initial values

$$\begin{aligned} y(0) &= c_0 \\ y_1(0) &= 2c_2 + 6c_3 \\ y_2(0) &= c_1 + c_2 + c_3 \end{aligned}$$

It is clear from the formulation that only $3N$ additions are needed (in addition to the computations of several constants). It is also clear that the polynomial evaluation can be implemented with a pipelined three-adder hardware so to produce one $y_3(x)$ per clock cycle.

**Bi-Variate Polynomial Evaluation** For any specific image row, the bi-variate polynomial can be reduced to a single-variable polynomial since the variable v has a given value. More specifically,

$$\begin{aligned} x(u,v) &= a_0 + a_1u + a_2v + a_3u^2 + a_4uv + a_5v^2 + a_6u^3 + a_7u^2v + a_8uv^2 + a_9v^3 \\ &= (a_0 + a_2v + a_5v^2 + a_9v^3) + (a_1 + a_4v + a_8v^2)u + (a_3 + a7v)u^2 + a_6u^3 \end{aligned}$$

Therefore, given a specific value of $v$, the coefficients of $x(u,v)$ as a polynomial in $u$ can be computed. It then takes 3N additions for the evaluation of the single-variable polynomial.

**Numerical Stability Considerations** The coefficients of the bi-variate polynomials have very different dynamic ranges. For example, the $a_0$ value may be from -5 to 5 while the $a_9$ value may be from -0.0000001 to 0.0000001. This is because the value of $x(u,v)$ (or $y(u,v)$) is bounded for image alignment and one of the contributing factor for $x(u,v)$ is from the multiplication of $a_9$ with $v^3$ which can be very large. From implementation point of view, the different dynamic ranges cause numerical problem due to finite precision. One way to ease the problem is to normalize the coefficients as follows. The technique is illustrated below with a single-variable polynomial.

Let $y(x) = c_0 + c_1x + c_2x^2 + c_3x^3$, where $x$ is in the range, 0, 1, 2, ..., N-1. The value of $c_3$ may be too small because (N-1)$^3$ is pretty large. To balance the ranges, $y(x)$ can be reformulated as

$$y(x) = c_0 + (c_1S)\frac{x}{S} + (c_2S^2)(\frac{x}{S})^2 + (c_3S^3)(\frac{x}{S})^3$$

Let $x_n = x/S$ and $g(x_n) = y(x)$, then

$$\begin{aligned} g(x_n) &= c_0 + (c_1 S)x_n + (c_2 S^2)x_n^2 + (c_3 S^3)x_n^3 \\ &= cn_0 + cn_1 x_n + cn_2 x_n^2 + cn_3 x_n^3 \end{aligned}$$

where $S$ is a constant whose value is chosen to be close to N and $g(x_n)$ becomes a polynomial whose coefficients are normalized to similar dynamic ranges. Note that $x_n$ is in the range $0, 1/S, 2/S, ..., (N-1)/S$. This requires a simple change to the recursive formulation. It can be shown that the previous evaluation formula become as follows.

$$\begin{aligned} y_3(x + \frac{1}{S}) &= y_3(x) + \frac{y_2(x)}{S} \\ y_2(x + \frac{1}{S}) &= y_2(x) + \frac{y_1(x)}{S} \\ y_1(x + \frac{1}{S}) &= y_1(x) + c \end{aligned}$$

where $c$ is a constant and the initial values can be computed from the normalized coefficients. Therefore the cost of normalization to the polynomial evaluation is simply some data shifting when $S$ is chosen to be as a power of two number. In the project, 512 is used for $S$.

## 2.3 Timing Requirements

The video input of the project has 30 frames a second and the frame resolution is 480 × 640 (height, width). As a result, there are 9.216 MegaBytes (= 640 × 480 × 30) of pixels per second for each channel of video signal. Since a video signal contains some horizontal/vertical blanking periods, the actual pixel rate is even higher. In fact, a clock rate of 12.5 MHz is necessary for the LIMAR hardware system for real time processing. This translates into 80 nsec of clock period during which time, for each channel, two third-degree bi-variate polynomials need to be evaluated and a bilinear interpolation needs to be calculated.

# 3 LIMAR Processing System

The current design of the LIMAR processing system is shown in Figure 2. The system processes and combines two channels of camera signals. The system contains two cameras, two monitors, two video I/O boards, a LIMAR processing board, two image alignment boards, a microcontroller board, and a personal computer (PC). Each video I/O board processes one channel of analog camera signal and converts it to a digital data string. The data buffering and video signal synchronization are also performed by the video I/O boards. Each channel of digital data is then sent to a corresponding image alignment board which, in real-time, aligns an image frame with an "ideal" image frame, or equivalently, removes some image spatial distortions. The aligned images are fed to the LIMAR processing board for further processing. The results are sent to video I/O boards and displayed on the monitors. A microcontroller board serves as the PC command interpreter and controls the whole LIMAR system. An application program with graphic user interface (under Windows 3.1 environment) runs on a PC and talks to the system through a PC parallel port. The usage of the PC parallel port increases the portability of the whole system.

In this project the system has been fully defined and several hardware boards have been implemented and tested so to verify the feasibility. Accomplished include (1) the definition of the LIMAR system bus, (2) the design, implementation, and testing of a video I/O board, (3) the design, implementation, and testing of an image alignment board, (4) the design, implementation, and testing of the microcontroller board, and (5) the implementation of the LIMAR control program.

## 3.1 LIMAR System Bus Definition

Since the LIMAR system bus is used by all the boards involved, a clear definition of its signals helps to decouple the designs of individual boards. The 16-bit PC/AT-bus is adopted and modified to serve as the system bus because of its low cost and popularity. By doing so, any AT-bus prototype board can be used for the LIMAR development. In fact, each of the board in Figure 2 uses a full size (13" x 4.5") AT-bus daughter board. The AT-bus connector is a 62-pin edge connector with a secondary 36-pin edge connector. The bus pinouts were re-defined for LIMAR system so that all the real-time signals are embedded in the system bus while some slower signals are implemented

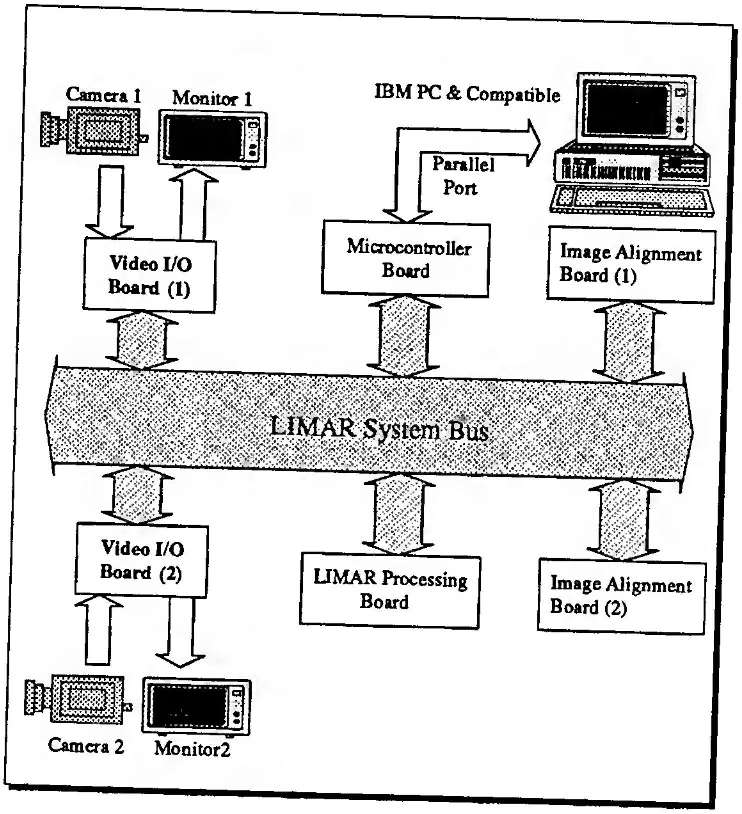


Figure 2: LIMAR Processing System Overview

with inter-board cable connections.

## 3.2 Video I/O Board

The video I/O board is designed to satisfy several requirements. First, it allows the PC to grab a single frame of image. Second, it allows the PC to display a single frame of image on a monitor. Third, it converts interlaced analog NTSC signal to non-interlaced digital signal and converts the digital signal back to analog in real-time. Lastly, it allows the PC to place a testing image for alignment. The last requirement makes it a lot more easier to test both the video board and the alignment board. The video I/O board as shown in Figure 3 contains the following components.

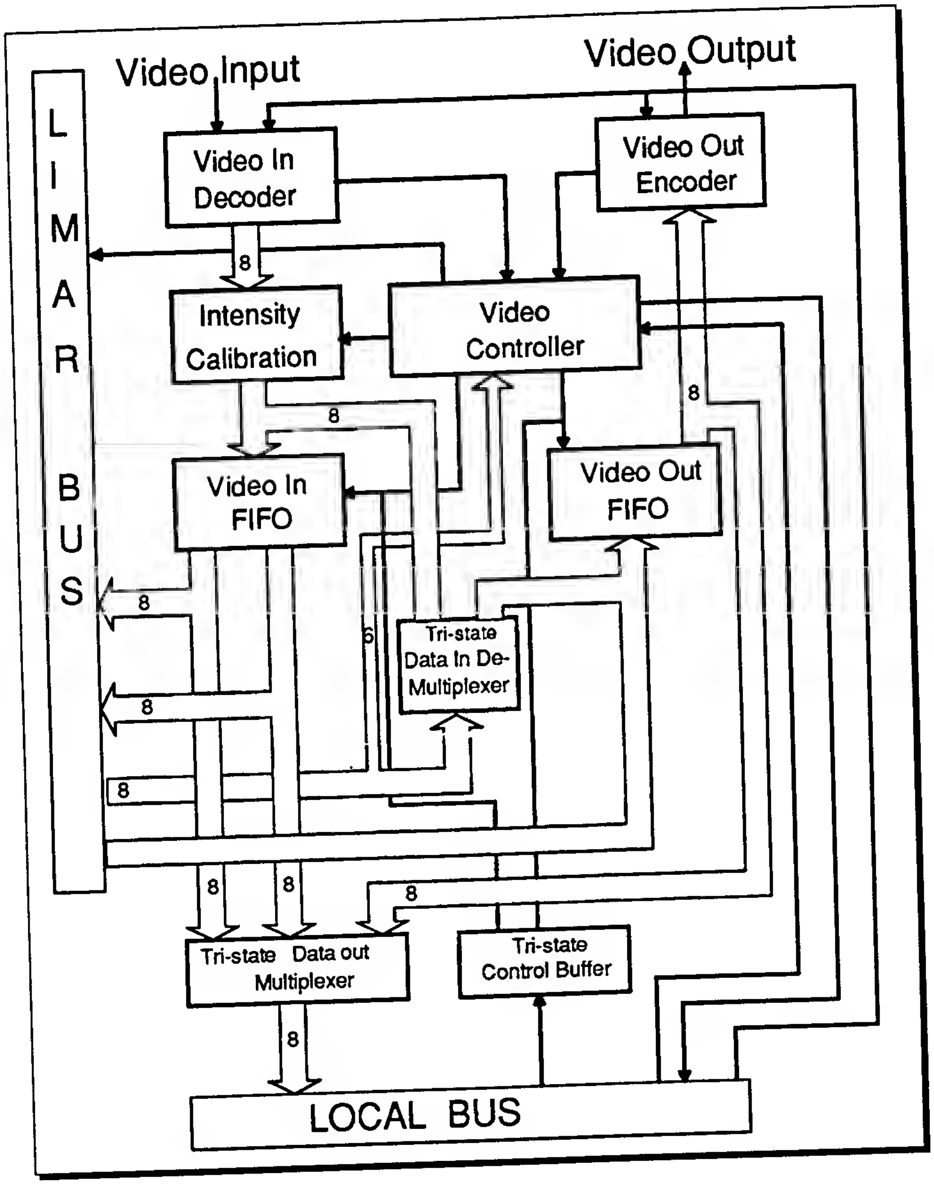


Figure 3: Video I/O Board

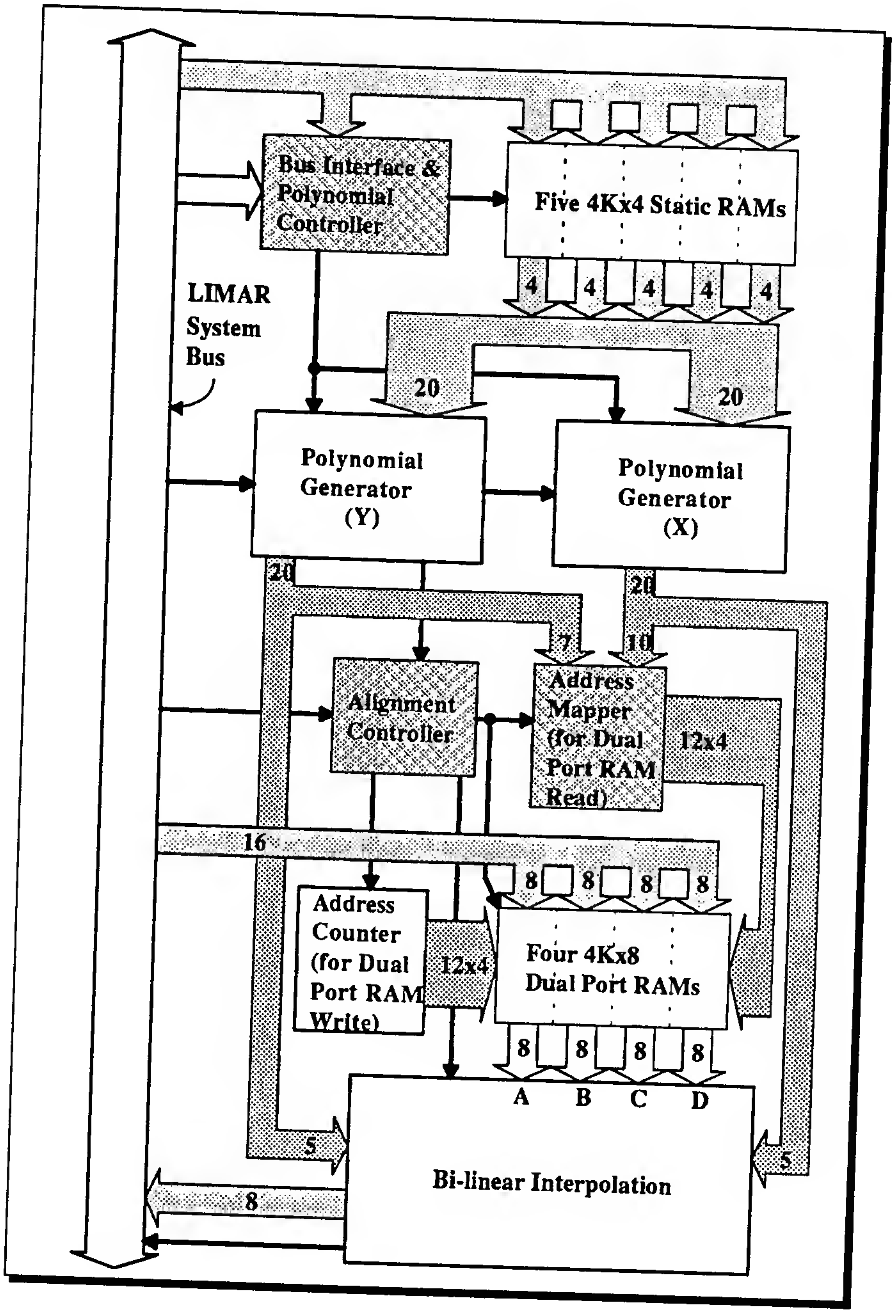


Figure 4: Alignment Board

1. Video In Decoder and Video In FIFO (First-In-First-Out): The decoder unit converts an analog interlaced NTSC video signal into a frame of digital non-interlaced data which is then stored into the input FIFO unit. The problem of synchronizing two independent cameras is solved with suitable timing control. The intensity calibration unit is currently bypassed since it is useful only when both channels are in place.

2. Video Out FIFO: The image data in the input FIFO is sent to the dual-port memory on the alignment board. For a complete system, the output of the alignment board is sent to the LIMAR board which collects and processes two channels of data. The output FIFO unit receives data from the LIMAR board. All the inter-board data transfers here are in real-time and go through the LIMAR system bus. Currently, the LIMAR board is not available yet. So an extra board is used to relay the output of the alignment board to the video out FIFO unit on the video board. The input to the FIFO is non-interlaced while the output of the FIFO is interlaced.

3. Video Out Encoder: The encoder unit converts the digital images from the output FIFO unit into NTSC signals which can be displayed on a monitor.

4. Video Controller: The controller unit is used to generate timing signals for the two FIFO units and the intensity calibration unit. It solves the camera synchronization problem and the interlaced image conversion problem. The controller is designed and implemented with an Altera Erasable Programmable Logic Device (EPLD) EMP5128 EPLD chip.

5. Tri-state Data Multiplexer, Demultiplexer, and Buffers: They are used to select and buffer desired data or control signals from/to bus. Note that a local bus is used between the video board and the microcontroller board to accommodate non-real-time signals.

## 3.3 Image Alignment Board

The image alignment is to convert an input image to another image according to two bi-variate polynomials, x(u,v) and y(u,v). An input pixel at location (u, v) is mapped to location (x, y) on an output image. If (x, y) is not a valid image location and falls among several pixels, a bi-linear interpolation is performed. The design of the image alignment board is sketched in Figure 4.

The design contains dual-port memory components to buffer data, polynomial generators to compute x(u,v) and y(u,v), a circuit to perform bi-linear interpolation, and two controllers. The polynomial generators and the circuit for the bi-linear interpolation are implemented with XILINX Field Programmable Gate Array (FPGA) XC4003 and XC4005, respectively. The two controllers are implemented with ALTERA EPLD EPM5000 series chips. The usage of these advanced programmable logic devices reduces chip counts and facilitates rapid implementation.

### 3.4 The Microcontroller Board

An MC6811 microcontroller receives commands from a PC program through the standard PC parallel port and controls the board accordingly. Since the standard PC parallel port is usually used as an output printer port, some effort was spent in converting it to a bi-directional port. The microcontroller sends $I^2C$ signals through local bus to control the video in decoder and video out encoder on the video I/O board. In addition, the microcontroller controls various data buffers (74244s) to select system working modes such as grabbing image, displaying image, or real-time processing.

### 3.5 The LIMAR Control Program

The LIMAR control program is an IBM-PC/Microsoft Windows-based program that serves to initialize and control the activities of the various components of the LIMAR system. Due to the speed required to process thirty 640x480 pixel images every second, the LIMAR control program does not take part in the real time operation of the system. Rather, the program will load startup parameters to various modules of the system and then send control signals to enable and disable those modules. The program has very friendly user interface.

## 4 Conclusion

The LIMAR device, which is conceptually the fastest image ranging device, utilizes two cameras to grab images which are required to be in full registration. In this project, camera registration algorithms were refined and incorporated into a customized processor design which can convert the image pair into range and intensity images in real-time. The detailed design of the processing unit and the control unit have been completed. Several hardware boards were implemented and tested to verify the design. Compared to the prototype LIMAR system which was assembled in 1992 and could not perform real-time computations due to the usage of a general purpose computer, the proposed processor represents a significant enhancement to the future LIMAR development program at the Avionic Directorate of Wright Laboratory.

# APPLICATIONS OF WAVELET SUBBAND DECOMPOSITION IN ADAPTIVE ARRAYS

Ismail Jouny
Assistant Professor
Department of Electrical Engineering

Lafayette College
Markle Hall, High Street
Easton, PA, 18042

Final Report for:
Research Initiation Program
Wright Patterson Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Lafayette College

December 1993

# APPLICATIONS OF WAVELET SUBBAND DECOMPOSITION IN ADAPTIVE ARRAYS

Ismail Jouny
Assistant Professor
Department of Electrical Engineering
Lafayette College

## Abstract

Pre-processing radar signals incident on an adaptive array by applying an invertible transformation such as wavelets is the focus of this study. The effect of wavelet subband decomposition of radar signals prior to adaptation using an LMS algorithm or an Applebaum processor on the adaptation rate of these processors is examined. The impact of wavelet transform on the bandwidth performance of adaptive arrays is also investigated. The performance of wavelet transform based array processors is compared with that of the FFT, and Cosine transform. The dynamic range of the array weights before and after wavelet transformation is also being examined. Simulations involving experimental radar data and different types of wavelets are also presented.

# APPLICATIONS OF WAVELET SUBBAND DECOMPOSITION IN ADAPTIVE ARRAYS

Ismail Jouny

## I. Introduction

Wavelet subband decomposition is a recently developed and rapidly evolving signal processing technology with emerging applications in speech compression, image coding, geophysics, radar and sonar signal processing.

The author, in a study conducted at Wright Patterson Air Force Base during the summer of 1992, used wavelet based target scattering features in radar target recognition. The results indicated that wavelet decomposition of radar cross section measurements (RCS) of unknown radar targets may be reliably used for target identification under different noise scenarios and without complete knowledge of the target azimuth position.

The following study is the result of a research effort supported by AFOSR research initiation program which followed the author's Summer Faculty Fellowship at Wright Patterson AFB.

This study focuses on the application of wavelet decomposition in adaptive antenna arrays, an idea that was developed during the summer of 1992 through conversations with WPAFB fellow researchers and senior scientists. The following summary of results shows that wavelets present a unique opportunity for improving the performance of adaptive antenna arrays.

Two specific points are addressed in this study, first the effect of wavelet decomposition on adaptation speed and convergence rate of adaptive systems including changes in the eigen structure of the covariance matrix. A connection between wavelet domain adaptation and other transform-domain adaptive processing techniques. Secondly, this study examines the effect of wavelet transform on the bandwidth performance of adaptive arrays. Scenarios of wideband jamming are simulated and the array signal-to-(noise+interference) ratio is examined. The study shows that improvement can be achieved regarding the adaptation and convergence speed of adaptive arrays as well as computational speed, but wavelet subband decomposition has little effect on the bandwidth performance of an array.

This research focuses on incorporating wavelet subband decomposition into adaptive array processing of both narrowband and wideband radar signals. Radar signals can be decomposed using wavelets into orthogonal and almost decorrelated subbands. Such a decomposition is usually performed using Fast Fourier Transform (which is equivalent to using a bank of non-orthogonal bandpass filters) or using tapped delay line cancellers.

Because the concept of wavelet analysis was just recently developed, a brief review of wavelet subband decomposition including definitions and properties is presented in the following section. Sections III and IV detail the work done and present new results concerning the performance of adaptive antennas. Section V presents conclusions and suggestions for future work.

## II. Wavelet Subband Signal Decomposition

Wavelet signal approximation is a powerful signal processing technique based on subband decomposition using orthogonal Finite Impulse Response (FIR) filters. These filters are generated from the so-called wavelet functions. This framework of signal processing, often called "multi-resolution analysis", provides the means for signal decomposition into orthogonal octave bands so that every subband can be processed separately. An exact replica of the original signal can be reconstructed using a set of orthogonal octave band filters.

The wavelet transform of a signal $x(t)$ is by definition a convolution of $f(t)$ with a wavelet $\psi(t)$ dilated by a factor $a$

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t-b}{a}\right)\, dt \tag{1}$$

which can be expressed in the frequency domain as

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} X(\omega)\Psi(a\omega)e^{j\omega b}\, d\omega \tag{2}$$

Thus, the wavelet transform of $x(t)$ is equivalent to filtering $x(t)$ using the bandpass filter $\Psi(a\omega)$ whose bandwidth varies as a function of scale $a$. For $a = 2^j$, $j \in Z$ these filters represent octave band filters. Clearly, large scales correspond to narrow smoothing filters that present a global view of the signal $x(t)$ and small scales correspond to wideband filters that extract the details of $x(t)$ (high frequency components). The signal $x(t)$ can be recovered from its wavelet transform using

$$x(t) = \frac{1}{a^{3/2}} \int_a \int_b W_x(a,b)\psi\left(\frac{t-b}{a}\right)\, da\, db \tag{3}$$

assuming that

$$\int_\omega \frac{|\Psi(\omega)|^2}{|\omega|}\, d\omega \ < \infty \tag{4}$$

or $\int_t \psi(t)\, dt = 0$. Fourier transform and Fourier series approximations of $x(t)$ require many expansion coefficients associated with high frequency components to model transient signals and perform the necessary cancellation, thereby permitting the inclusion of high frequency noise. In contrast to that, wavelet analysis permits a selective mode representation of the signal due to the compact nature of the analyzing wavelet (limited duration of $\psi(t)$), and is therefore particularly suited for analyzing transient signals and singularities. When using Fourier transform, we expand the signal $x(t)$ using orthogonal complex sinusoidal functions. Similarly, with wavelets we expand the signal using dilated and translated version of a mother wavelet $\psi(t)$.

Orthogonality is an important element of wavelet analysis where a wavelet $\psi(t)$ is orthogonal to its own dilations $\psi(at)$ and translations $\psi(t-b)$. Orthonormal expansions are smooth and smooth functions have a rapidly decaying Fourier representation which enhances the frequency resolution attained using wavelet decomposition. Wavelet transform parameters can be discretized so that

$$c_{mn} = \frac{1}{a_0^{m/2}} \int_t x(t)\psi\left(\frac{t - na_0^m T}{a_0^m}\right) dt \tag{5}$$

where $a = a_0^m$ and $b = na_0^m T$, and $T$ is the sampling period. The signal $x(t)$ can then be recovered from its expansion coefficients using

$$x(t) = A\sum_m \sum_n c_{mn}\psi\left(\frac{t - na_0^m T}{a_0^m}\right) \tag{6}$$

where $A$ is a constant. Orthogonality in this case is equivalent to

$$\int_t \psi(t)\psi\left(\frac{t-n}{a_0^m}\right) = 0, \quad \forall m,n - \{0,0\} \tag{7}$$

The accuracy of the reconstruction depends on the adopted wavelet basis and whether it constitutes a tight Frame. The case where $a_0 = 2$ is known as the dyadic wavelet transform.

The above definition of the wavelet transform is for continuous signals. For discrete signals, wavelet transform is implemented using a bank of bandpass and lowpass discrete time filters that can be reconstructed using few coefficients. The filters needed have orthogonal impulse responses that can be derived using simple recursion formulae.

The wavelet transform of a discrete time sequence $x(k)$ is essentially a multiresolution characterization of $x(k)$. Wavelet decomposition of $x(k)$ is represented by a set of detail signals that are associated with the high frequency components of $x(k)$

and a final coarse approximation. Mallat [17] has developed a very efficient multiresolution wavelet decomposition algorithm that limits the number of wavelet expansion parameters to $N$ where $N$ is the length of the data sequence $x(k)$. As the signal $x(k)$ propagates through the filter bank tree of lowpass and highpass filters, the output of the highpass filter $G(z)$ at stage $m$ is a sampled version of the wavelet transform of $x(k)$ at scale $2^m$. At each stage, the bandwidth of both filters is halved with the high halfband associated with the highpass filter and the low halfband associated with the lowpass filter. The dyadic discrete wavelet transform is essentially a decomposition of the spectrum of $x(k)$, $X(e^{j\omega})$ into orthogonal subbands defined by

$$\frac{1}{2^j T} \leq \omega \leq \frac{1}{2^{j+1} T} \quad ; \; j = 1, \ldots, J \tag{8}$$

where $T$ is the sampling period associated with $x(k)$. Therefore, wavelets are unique in offering a framework for examining radar signals at different resolutions (different frequency bands) and processing each component separately.

## III. Wavelets and Adaptive Arrays

Adaptive array processing with applications in radar and communications is a discipline that has received considerable attention in the last few decades. There are numerous studies addressing almost every aspect of the problem of adaptive signal processing. Rejection of intentional jamming and scattered interference is one of the many applications of adaptive array processing. Decorrelation of signal components, for the purpose of simplifying the adaptation procedures, by means of subband decomposition either using the FFT or using tapped delay-line cancellers has also been a subject of great significance in adaptive processing.

The bandwidth performance (or nulling bandwidth) is an important factor in designing adaptive arrays. To address this problem, researchers have proposed FFT processing as a tool for band partitioning the frequency response of the received signal and adapting each band separately. Others have shown that a transversal filter constructed as a tapped delay-line does improve the bandwidth performance of adaptive arrays. L. E. Brennan compared the performance of both FFT based processing and transversal filters in improving the cancellation ratio of sidelobe cancellers assuming mismatched receiver characteristics. Mismatch between receivers in different channels could be simulated as random pole placement, or shift in the center frequency and a difference in the bandwidth of the receiver. Different receiver mismatch scenarios were considered in Brennan's study and the transversal filter method constantly outperformed the FFT approach. Later, Compton showed that both methods have equivalent bandwidth performance (improving nulling bandwidth) provided that the delay between taps is identical to the delay between samples of the FFT. He also showed that no invertible transformation can be inserted between the delay-line taps and the weights that may improve the nulling bandwidth of the array. However using the FFT technique reduces the correlation between samples in disjoint frequency subbands which leads to a block diagonal covariance matrix. Thus with FFT processing the weights are computed and adapted separately in each band. Although, samples in different frequency bands are usually not completely decorrelated and the covariance matrix is not absolutely block diagonal, frequency domain adaptive filtering reduces the eigenvalue spread of the data autocorrelation matrix.

## 1. Array Structure

The adaptive arrays examined in this report are known as the Applebaum array and the LMS array. Although, we focus on the LMS adaptation algorithm, the results can be readily generalized to include the Applebaum array.

An N elements array is shown in Figure 1 where $X$ denotes the input vector

$$X = \begin{bmatrix} X_{11} \\ X_{12} \\ \dots \\ X_{1K} \\ X_{21} \\ \dots \\ X_{2K} \\ \dots \\ \dots \\ X_{NK} \end{bmatrix} \quad W = \begin{bmatrix} W_{11} \\ W_{12} \\ \dots \\ W_{1K} \\ W_{21} \\ \dots \\ W_{2K} \\ \dots \\ \dots \\ W_{Nk} \end{bmatrix} . \tag{9}$$

The array includes $K$ tap delay elements with $NK$ weights, henceforth the weight vector $W$. The weights are controlled using either the LMS algorithm (where a reference signal is required) or the Howell-Applebaum algorithm where both algorithms yield the same optimal solution. The output signal is denoted by $s_X$ where $s_X = X^T W$ ($T$ denoting transpose). The optimal weight solution for both algorithms is

$$W_{opt} = \Phi_x^{-1} S_X \tag{10}$$

where $S_X$ is the steering vector ($S_X = E\{Xr(t)\}$, $r(t)$ being the reference signal) . The matrix $\Phi_X$ is the covariance matrix of the input vector $X$, and defined as

$$\Phi = \begin{bmatrix} E\{X_{11}X_{11}\} & \dots & E\{X_{11}X_{1K} & \dots & \dots & \dots & \dots & \dots & E\{X_{11}X_{NK}\} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ E\{X_{21}X_{11}\} & \dots & E\{X_{21}X_{1K}\} & \dots & \dots & \dots & \dots & \dots & E\{X_{21}X_{NK}\} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ E\{X_{Nk}X_{11} & \dots & E\{X_{Nk}X_{1K}\} & \dots & \dots & \dots & \dots & \dots & E\{X_{NK}X_{NK}\} \end{bmatrix} \tag{11}$$

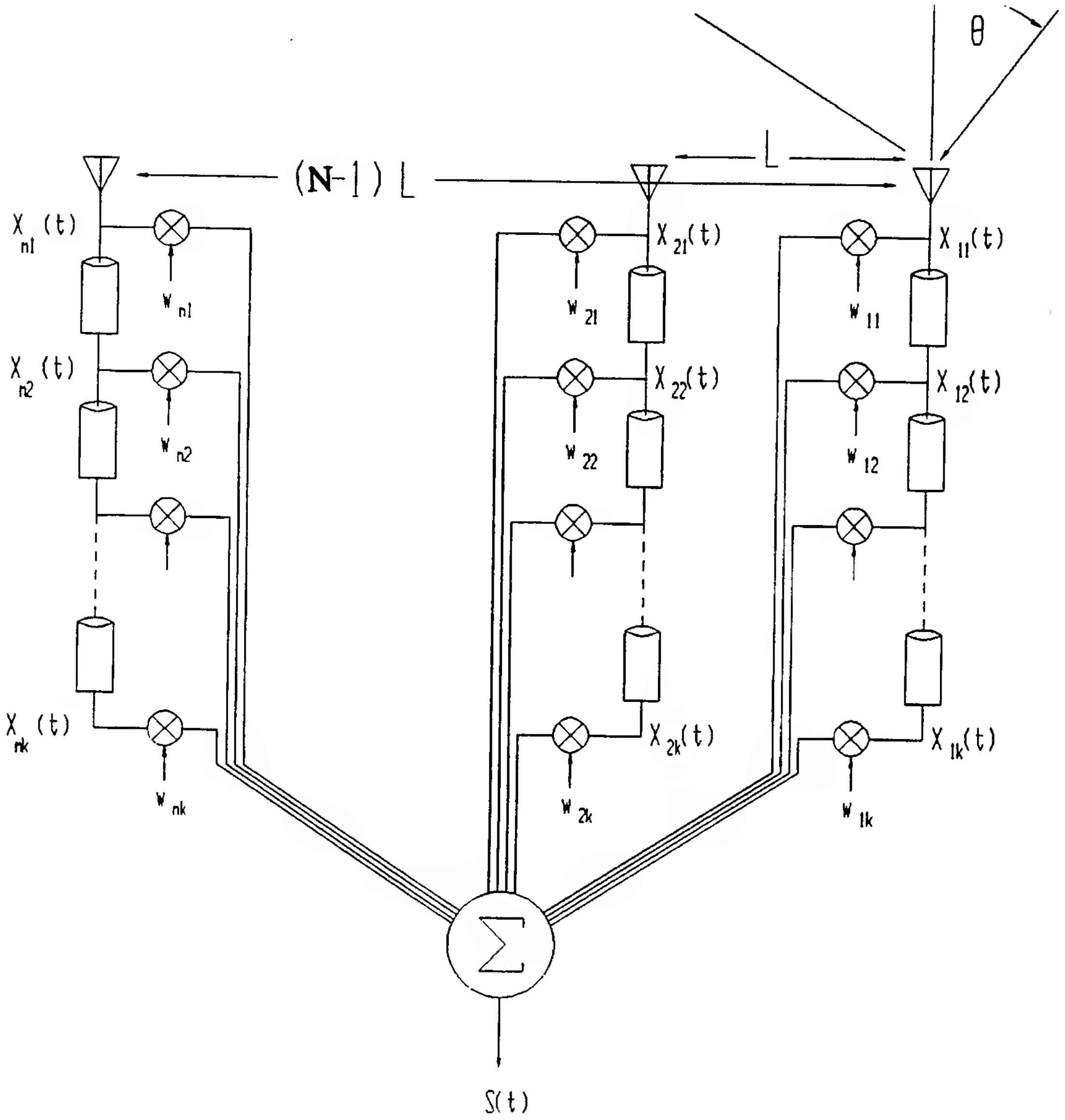


Figure 1: Adaptive antenna array of $N$ elements each with K tap-delays

where $E\{.\}$ denotes the expected value. Each weight vector is adjusted according to the rule

$$\frac{dW}{dt} = -k \, \nabla_W \, E\{[r(t) - s(t)]^2\} \tag{12}$$

where $r(t)$ is the reference signal (correlated with the desired signal $d(t)$). If the input is discrete then the above adaptation rule at time $n$ reduces to

$$W_{ij}^{n+1} = W_{ij}^{n} + 2\mu X_{ij}^{n} \, [r(n) - s(n)] \tag{13}$$

To ensure stability of the LMS algorithm, $\mu$ can be chosen as

$$0 < \mu < \frac{2}{NKE\{X^2\}} \tag{14}$$

In fact for the weights to converge, the adaptation parameter $\mu$ should be chosen such that

$$0 < \mu < \frac{2}{tr\,(\Phi_X)} \tag{15}$$

Notice that by taking the expected value of both sides of the adaptation equation (written in vector form hereafter)

$$E\{W^{n+1}\} = W^n + 2\mu E\{X \left[r(n) - X^T W\right]\} \tag{16}$$

which can be rewritten as

$$E\{W^{n+1}\} = (I - 2\mu\Phi)\, W + 2\mu E\{X r(n)\} \tag{17}$$

where $\Phi = E\{XX^T\}$. The above equation can be further simplified to yield

$$E\{W^{n+1} - W_{opt}\} = (I - 2\mu\Phi)^n \, E\{W^0 - W_{opt}\} \tag{18}$$

By properly choosing $\mu$, the above equation leads to (10).

## 2. Transform-Domain Adaptation

The convergence of the above adaptation algorithm is dependent on the eigen-spread (assuming stationary input) $\lambda_{max}/\lambda_{min}$ which is known as a the *condition number* which is an indicator of the rate of convergence of the adaptive algorithm and provides diagnostic description of the ill-conditioning of the matrix $\Phi$. Therefore, in order to achieve high convergence rates (which is a crucial factor in radar technology), the adaptation must be performed in an orthogonal domain obtained by transforming the input vector $X$. The Karhunen-Loeve transform would be the ideal transform which produces completely orthogonal signal components but computing the KLT is very dependent on the exact estimate of $\Phi$ which is complicated and impractical. Two other possible options which are not based on transformation but can improve the adaptation speed are; the recursive least square algorithm (RLS) and the Gram-Schmidt orthgonalization method. These options are tedious and computationally demanding.

Alternatively we may use the following transforms:

1) Discrete Fourier transform (implemented with the FFT) where the input signal $X$ is transformed into

$$\begin{aligned} Y_{ik} &= \frac{1}{\sqrt{K}} \sum_{n=0}^{k-1} X_{in} e^{\frac{-j2\pi nk}{N}} \\ X_{in} &= \frac{1}{\sqrt{K}} \sum_{k=0}^{K-1} Y_{ik} e^{\frac{j2\pi nk}{N}} \end{aligned} \qquad (19)$$

2) the Cosine transform

$$Z_{i0} = \frac{\sqrt{2}}{K} \sum_{n=0}^{K-1} X_{in} \qquad (20)$$

$$Z_{ik} = \frac{2}{N} \sum_{n=0}^{K-1} X_{in} \cos\left(\frac{\pi(2n+1)k}{2N}\right) \qquad (21)$$

which can be implemented using a bank of bandpass filters.

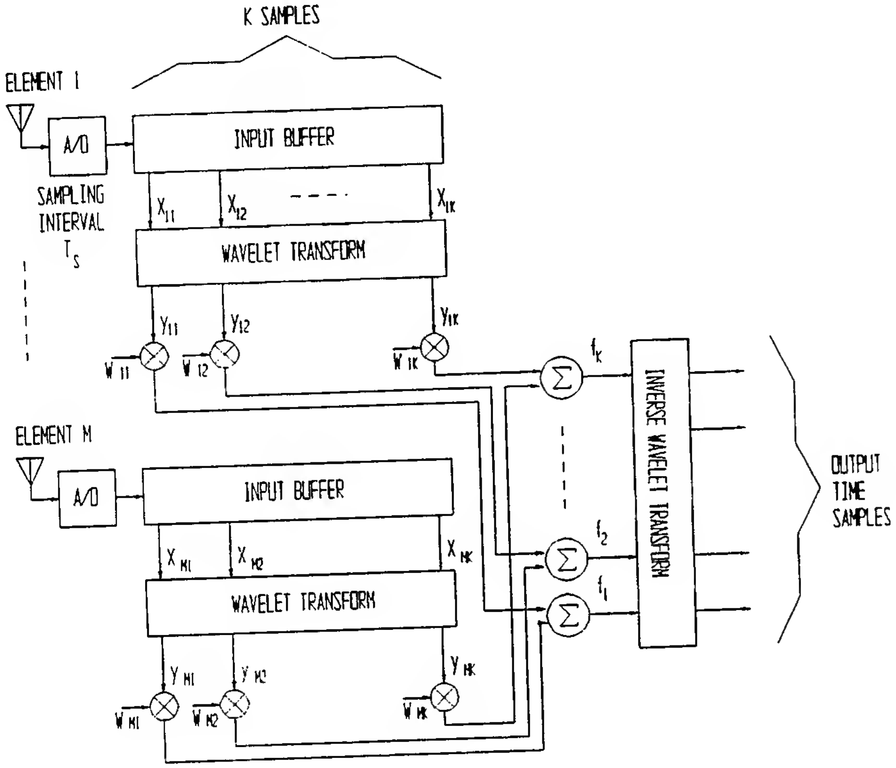


Figure 2: An M elements adaptive array with wavelet transform as a pre-processor

3)Wavelet transform (defined in Section II and implemented using a cascade of low pass and high pass filters) as shown in Figure 2. The wavelet transform at resolution $M$ [3] of the discrete signal $X_i$ (where $X_i$ is of length $K$ and entering the $i-th$ array element) can be expressed as

$$V = Q_{log_M K}(\ldots(Q_2(Q_1 X_i))\ldots) \tag{22}$$

$$Q_j = \begin{bmatrix} I_{K-M^{log_M K-j+1}} & \mathbf{0} \\ \mathbf{0} & D_j \end{bmatrix} \tag{23}$$

where $D_j$ is the wavelet analysis matrix at stage $j$, and $I$ is the identity matrix.

The transform domain adaptive antenna array is shown in Figure 2 where the input vector $X$ is transformed into an output vector $Y$. Notice that in all of the above transforms, the output could be written as $Y = \Gamma X$, where $\Gamma$ represents the transformation matrix of rank $N$. The new weight vector $U$ is adapted using

$$U^{n+1} = U^n + 2\mu Y\left[r(n) - s_Y\right] \tag{24}$$

where $s_Y = Y^T U$ and $\mu$ is normalized with respect to $E\{Y^T Y\}$. If $\mu$ is properly chosen, then the optimal weight solution is

$$U_{opt} = \Phi_Y^{-1} S_Y \tag{25}$$

where $\Phi_Y = E\{YY^T\} = E\{\Gamma X X^T \Gamma^T\}$ and $S_Y = E\{Yr(n)\}$. Therefore,

$$\Phi_Y = \Gamma \Phi_X \Gamma^T \tag{26}$$

thus $U_{opt} = \Phi_Y^{-1} S_Y = |Gamma^{-T}\Phi_X^{-1}\Gamma^{-1}E\{\Gamma X r(n)\}$ can expressed as and the statement $S_Y = Y^T U = \Gamma X^T W = \Gamma S_X$ is thus equivalent to

$$\begin{aligned} U_{opt} &= \Gamma^{-T}\Phi_x^{-1}\Gamma^{-1}\Gamma S_X \\ &= \Gamma^{-T} W_{opt} \end{aligned} \tag{27}$$

It is shown in [1] that the condition number of $\Phi_Y$ is always less than that of $\Phi_X$ which implies that adaptation in the transform domain will proceed faster than that in the time domain.

The performance of wavelet based LMS filters depends on the convergence parameter and the type of mother wavelet used. In [5], the wavelet packet approach is used to improve the rate of convergence of adaptive arrays. The method used in [5] is based on maximizing the cross-correlation between received signals, and the decomposition scheme is chosen so that wavelet processed signals are maximally correlated in each of the subbands. This approach may produce better convergence rates than direct wavelet subband decomposition, but requires a nontrivial additional computational burden which makes the proposed method even more demanding than the recursive least square approach or the Gram-Schmidt method.

The question that arises immediately after employing wavelet transform in adaptive antenna arrays is whether the covariance matrix of the transformed signal is totally diagonal (i.e. its condition number is one). Neither the FFT, nor the Cosine transform produce totally decorrelated signal components or a diagonal covariance matrix because of limitations concerning the implementation of these transforms. The wavelet transform, which is also an orthogonal subband decomposition scheme, does not produce completely decorrelated signals either. In fact it is shown in [3] that, for a large class of random processes, off the diagonal elements of the covariance matrix can be generally expressed in the following form

$$\Phi_x(i,j) = i^{\alpha(j)} e^{-i\beta(j)} \left(a_0(i) + o(1)\right) + a_1(i) \tag{28}$$

and are, despite decreasing at a fast rate as a function of time, not identically zero. The correlation between signal components that belong to different subbands decays even faster than what is indicated in the above equation. Therefore the covariance matrix of wavelet transformed radar signal is near diagonal.

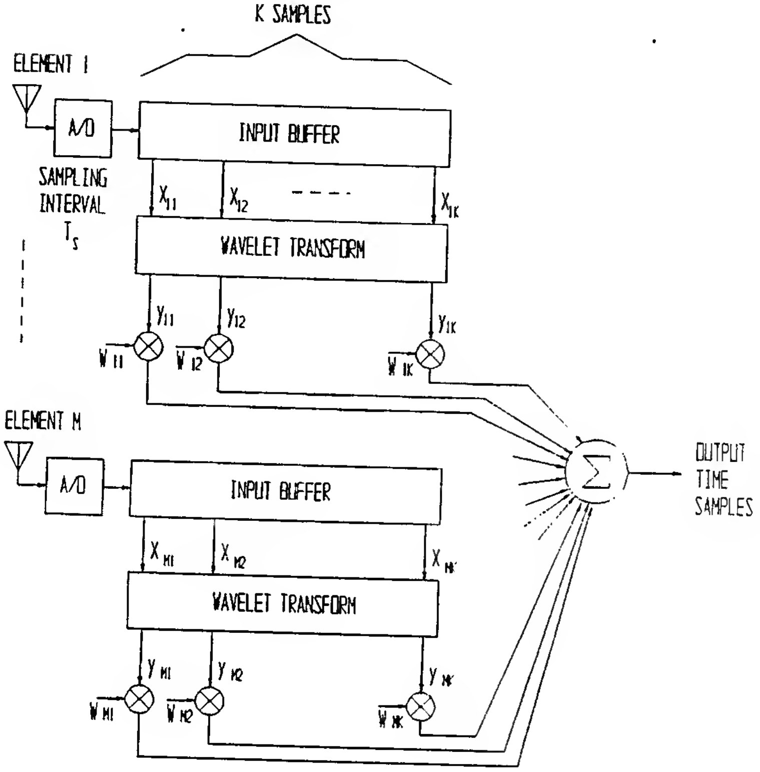


Figure 3: Adaptive array with adaptation in the wavelet transform domain (time-scale)

Erdol and Basbug [2] proposed an alternative incorporation of wavelet decomposition into adaptive arrays through sampling the direct wavelet transform of the incoming radar signal and truncating the dyadic wavelet series both in scale and in translation. This approach requires computational complexity in the order of the number of samples $NK$ but no truncation criterion is available and improper sampling of the wavelet transform may not yield the necessary improvement in adaptation rate. Alternatively, we propose to use the regular subband decomposition scheme which requires arithmetic operations directly proportional to the number of data samples $NK$ and thus compares favorably with that of the FFT or the Cosine transform from a computational standpoint.

## 3. Error Analysis

The minimum asymptotic error (Wiener solution) achieved with adaptation in the wavelet domain $\epsilon^w$ is related to that of the time domain $\epsilon_t$ as follows (using similar argument to [2])

$$\epsilon^w = \epsilon_t - S_Y^T \left(\Gamma^T \Phi_Y^{-1} \Gamma - \Phi_x^{-1}\right) S_X \tag{29}$$

thus the asymptotic error obtained in the transform domain could be lower than its counterpart in the time domain when $S_Y^T \left(\Gamma^T \Phi_Y^{-1} \Gamma - \Phi_X^{-1}\right) S_X$ is positive semi-definite. The steady state mean square error is defined as

$$\epsilon_{ss} = \epsilon + \epsilon_\Delta \tag{30}$$

where $\epsilon_\Delta$ is the excess mean square error and $\epsilon$ is the minimum error obtained by Weiner solution. The excess mean square error of the time domain [4] and transform domain LMS are

$$\epsilon_\Delta \;=\; \frac{1}{2}\mu Tr(\Phi_X)\epsilon_{min} \tag{31}$$

$$\epsilon_{\Delta}^{w} = \frac{1}{2}\mu Tr(\Phi_Y)\epsilon_{min}^{w} \tag{32}$$

where $Tr()$ denotes the trace of a matrix. Note that $Tr(\Phi_Y)$ is upper bounded by the energy of $Y$ which is upper bounded by the energy of $X$ (wavelet transform reserves energy). Therefore, the excess mean square error of the wavelet transform LMS is upper bounded by that of the time domain LMS.

## 4. The Adaptation Parameter $\mu$

The performance of the wavelet based adaptation scheme also depends on the choice of adaptation parameter $\mu$. Stability requires that

$$\mu \leq \frac{1}{\lambda_{\Phi_Y}^{max}} \tag{33}$$

where $\lambda_{\Phi_Y}^{max}$ is the largest eigenvalue of the covariance matrix $\Phi_Y$. Alternatively, $\mu$ could be chosen as a function of time $n$ where $\mu_n = \frac{\mu}{E\{Y_n^T Y\}}$. This approach improves the convergence rate of the array but depends on the fluctuations of signal power including steady state. A better convergence rate can be achieved if $\mu$ is dependent on the inverse of the covariance matrix.

$$W^{n+1} = W^n + 2\mu\Phi_X^{-1}X\left[r(n) - s_X\right] \tag{34}$$

$$U^{n+1} = U^n + 2\mu\Phi_Y^{-1}\left[r(n) - s_Y\right] \tag{35}$$

This approach [1, 2, 4] is known as self-orthogonalization LMS and results in faster convergence rate when applied in the time domain. The parameter $\mu$ in this case can be chosen as

$$0 < \mu < \frac{2}{NK}. \tag{36}$$

The eigenvalues of the self-orthgonalization matrix are all one, which results in significant improvement in adaptation speed. An alternative wavelet adaptation parameter that is exponentially weighted according to the subband of interest could also improve the adaptation rate but requires careful adjustment of the parameter $\mu$. Cholesky decomposition can also be employed to enhance the adaptation speed [3] of antenna arrays where the adaptation parameter $\mu$ is pre-conditioned by a non-diagonal matrix $\Lambda$ (obtained by solving the equation $\Lambda^2 f = S_Y$) where $f$ is a constraint vector and the weights are then updated using $W^{n+1} = W^n + 2(\mu/\Lambda^2)f$. This can be attempted with a modest increase in computational cost but is not included in the experimental phase of this study. To this end, our approach is based on computing the wavelet transform using a cascade of orthogonal low pass and band pass filters and using either a fixed $\mu$ or an adaptation parameter which is normalized with respect to signal energy.

## 5. Adaptation Rate of Wavelet Domain LMS ARRAYS

Radar cross section measurements of a DC10 aircraft model were used to examine the impact of transformation on the adaptation speed of an antenna array. The data is recorded in the frequency range 1-12 GHz with increments of 50 MHz and represent scattering in the resonance region. The eigenvalues of a ninth order covariance matrix are shown in Table 1. A lag of 9 was arbitrarily chosen for convenience, and Table 1 also shows the covariance elements of the DC10 RCS signal for lags 0 to 9. The condition number $\lambda_{max}/\lambda_{min}$ of the covariance matrix before and after transformation of the DC10 data is shown in Table 2 (in addition to $\lambda_{min}$ and $\lambda_{max}$). Clearly, the condition number of the wavelet transformed data is less than that of time domain, FFT, and cosine transform. Therefore, the covariance matrix of the wavelet transformed signal is less ill-conditioned than that of the FFTed or the Cosine transformed data.

The adaptation rate of the arrays shown in Figures 4 and 5 was examined using three types of signals; two sinusoids in noise, colored noise, and noisy radar data. The

Table 1: The covariance elements of a DC10 RCS data and the eigenvalues of the covariance matrix

| $\lambda$ | $\phi_X$ |
|---|---|
| 62.8 | 754 |
| 60.2 | 707 |
| 71.5 | 651 |
| 34.8 | 623 |
| 14.7 | 582 |
| 3 | 553 |
| 144.9 | 538 |
| 260.4 | 517 |
| 664.9 | 489 |
| 6.2 | 462 |

Table 2: Condition number of DC10 covariance matrix before and after transformation

| Transform | $\lambda_{min}$ | $\lambda_{max}$ | $\lambda_{max}/\lambda_{min}$ |
|---|---|---|---|
| Time Domain | 2.955 | 6230 | 2108 |
| FFT Domain | 139 | 11800 | 92.5 |
| DCT Domain | 0.0633 | 12.99 | 205.2 |
| Wavelet(D4) | 1.2 | 98.2 | 81.8 |

first two examples are commonly used in the literature to demonstrate new adaptive arrays algorithms, and the third example is associated with real radar data. The sinusoidal signal is given as

$$x(k) = 0.1\cos\left(\frac{\pi k}{15}\right) + \cos\left(\frac{5\pi k}{16}\right) + \text{noise}. \tag{37}$$

Figure 6 shows a comparison between the LMS errors of time domain, Cosine transformed, and wavelet transformed data. The wavelet used to generate Figure 6 is the Daubechies 4-tap wavelet (D4). Figure 6 shows that the rate of convergence of the wavelet transformed signal is higher than that of the DCT, FFT, or time-domain data. Wavelets such as the Daubechies 8 and 19 tap filters produced similar results to Figure 6. Figure 7 shows a comparison between the convergence rates of two transforms using a non-smooth signal (generated using colored noise and deterministic components) as input. The wavelet used to generate Figure 7 is the Daubechies D4 wavelet. The error is averaged over 50 iterations and $\mu = 0.005$. Figure 8 shows similar convergence curves when the input signal represents the RCS measurements of a DC10 model aircraft. Again, the adaptation of wavelet transformed data is faster than other forms of transformation. Figure 9 shows a comparison between the adaptation rate of FFTed data and that of wavelet transformed data using a wavelet of filter order 19 (D19) and the results are relatively similar to those shown in figures 6,7 and 8.

## 6. Wavelet Transform and Weight Dynamic Range

The weights dynamic range is another important issue of practical hardware significance in adaptive arrays. To examine the impact of wavelet transformation on the weight dynamic range let $e_{j\Gamma}$ be the eigenvectors of the transformation matrix $\Gamma$ then [39] the weight vector $W$ can be expressed as a linear sum of $e_{j\Gamma}$

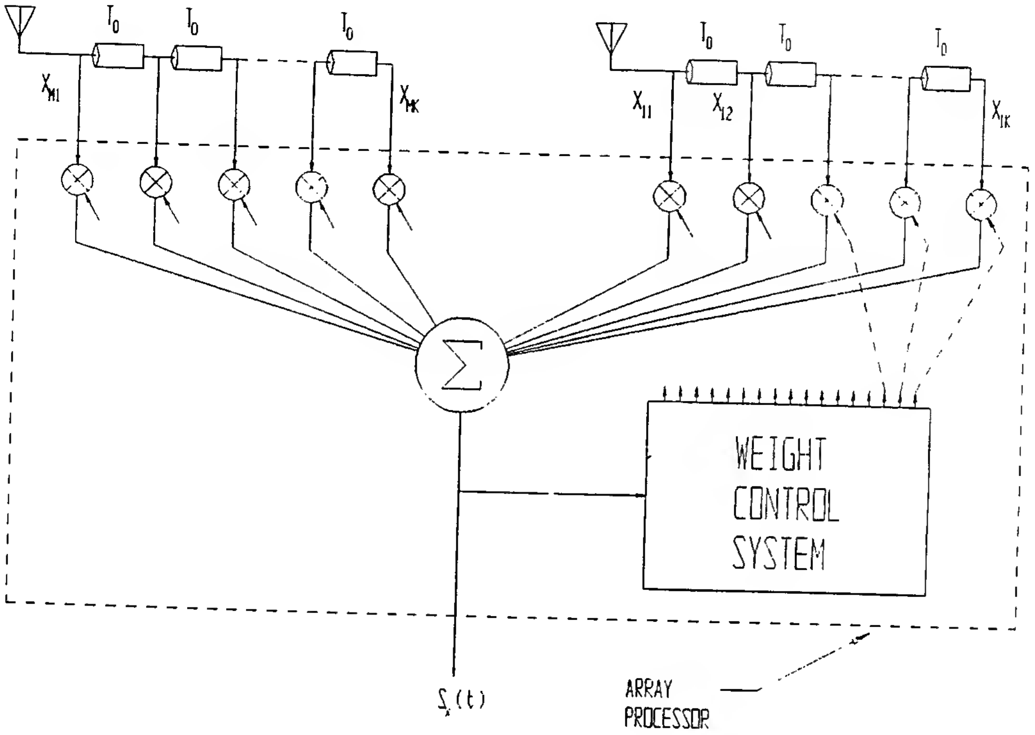


Figure 4: An LMS adaptive antenna array.

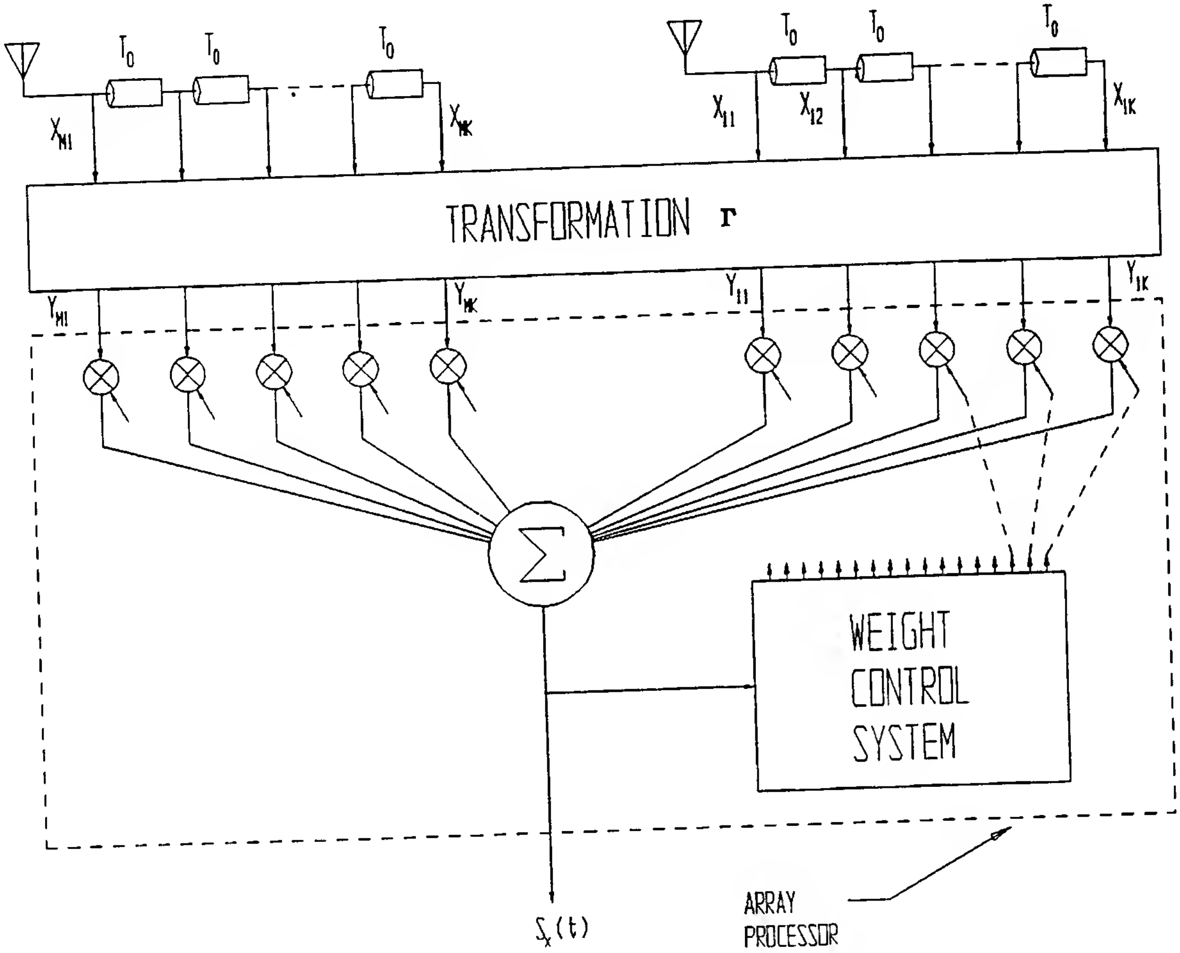


Figure 5: Wavelet transform domain LMS array.

$$W = \sum_{j=1}^{NK} \alpha_j e_{j\Gamma}, \tag{38}$$

hence, given that

$$\Gamma^{-1} = \frac{1}{\lambda_{j\Gamma} e_{j\Gamma}} e_{j\Gamma}^{*T} \tag{39}$$

where $\lambda_{j\Gamma}$ are the eigenvalues of $\Gamma$, the new weight vector $U = \Gamma^{-1}W$ can be expressed as

$$U = \sum_{j=1}^{NK} \frac{\alpha_j}{\lambda_{j\Gamma}} e_{j\Gamma} \tag{40}$$

Therefore, the elements of the new weight vector $W$ can be smaller than their counterparts in $W$ when all the eigenvalues of $\Gamma$ are greater than one ($\lambda_{j\Gamma} > 1$). Thus, the dynamic range of weight vector elements can be improved if $\lambda_{j\Gamma} > 1 \ \forall j$. This is the case with wavelet transform. For example, consider the Haar wavelet transform of 4 data points. The transformation matrix is

$$\Gamma = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \tag{41}$$

the eigenvalues of $\Gamma$ are 2.576, 2.576, 2.4 and 2.0 (all $> 1$). Therefore, the wavelet transform (see Figure 3) improves the weight dynamic range of an adaptive array.

## 7. Wavelet Transform and The Applebaum array

The above discussion about employing the wavelet transform applies to LMS based adaptive antenna arrays. Similar arguments apply to the Applebaum array. To prove this claim, we use arguments similar to those presented in [39]. Let $G$ be the steering vector and $\nu$ be a constant. Then, the optimum weight vector of the Applebaum array is given as

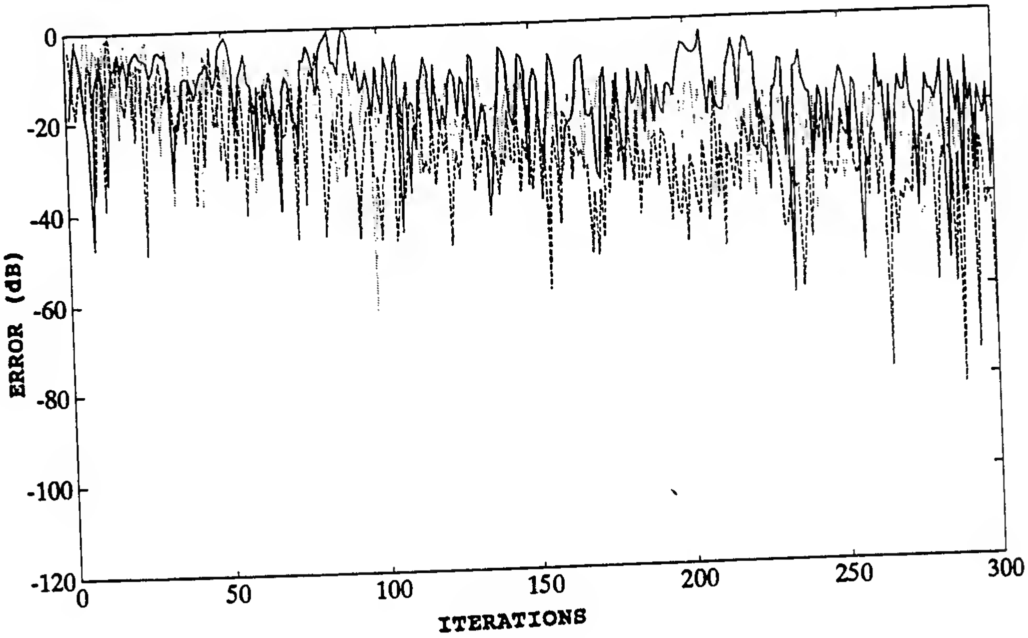


Figure 6: Comparison between adaptation rate of time-domain LMS (solid), Discrete Cosine Transform (DCT) LMS (dotted), and Wavelet (using D4) LMS with a sinusoidal input (dashed).

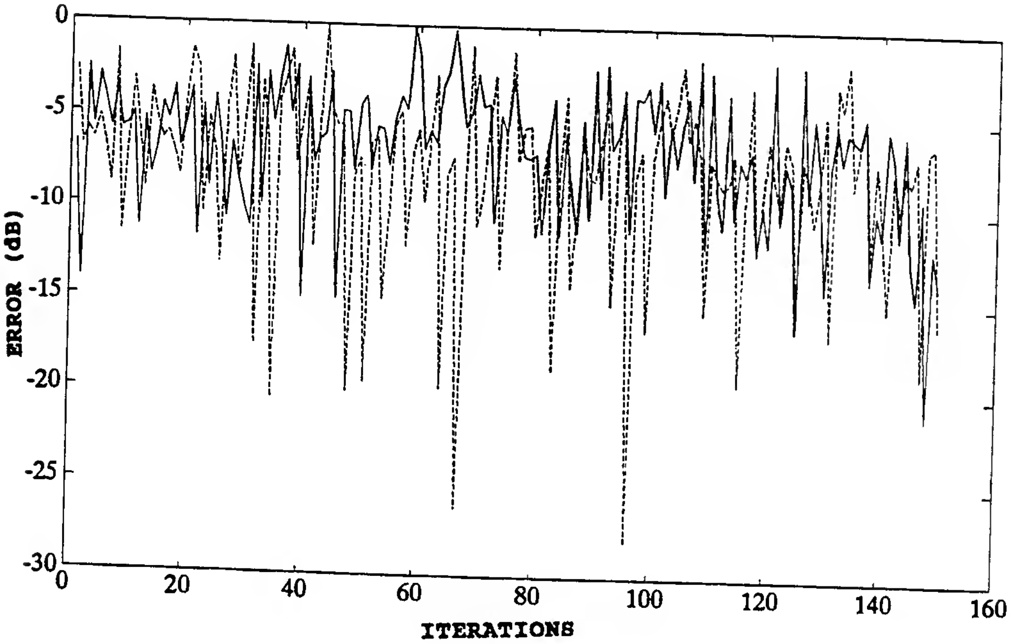


Figure 7: Adaptation rate for FFT LMS (solid) and Wavelet domain (D4) LMS (dotted) with a non-smooth signal input.

$$W_{opt} = \nu \Phi_x^{-1} G \tag{42}$$

and the array output $s_X$ is

$$s_X = X^T W = \nu X^T \Phi_X^{-1} G \tag{43}$$

Let $\Gamma$ be the transformation matrix as in the LMS case, where the incoming radar signal is being transformed prior to the Applebaum array processor, then the new optimum weight vector $U$ is

$$U_{opt} = \nu \Phi_Y^{-1} H \tag{44}$$

where $H$ is the new steering vector. Recall that $\Phi_Y = \Gamma \Phi_X \Gamma^T$ then

$$U_{opt} = \nu \Gamma^{-T} \Phi_X^{-1} \Gamma^{-1} H \tag{45}$$

and the output $s_Y = Y^T U$ is (recall $Y = \Gamma X$)

$$s_Y = \nu X^T \Gamma^T \Gamma^{-T} \Phi_X^{-1} \Gamma^{-1} H. \tag{46}$$

Clearly, if we chose $H = \Gamma G$ then $s_Y = s_X = \nu X^T \Phi_X^{-1} G$. Therefore, by incorporating a wavelet transform (or any transform $\Gamma$) prior to adaptation by an Applebaum processor, the output vector remains the same. The transform $\Gamma$ adjusts the adaptation speed of the array but does not change the signal-to-noise ratio of the array output.

## 8. Computational Requirements

Transformation of an incoming signal prior to noise cancellation or interference rejection with an array processor is a computationally demanding procedure and the cost depends on whether the signal is real or complex. If the signal is real then, for example, it is shown [4] that the number of multiplications needed is

$$N \;=\; 2NK + 5 \tag{47}$$

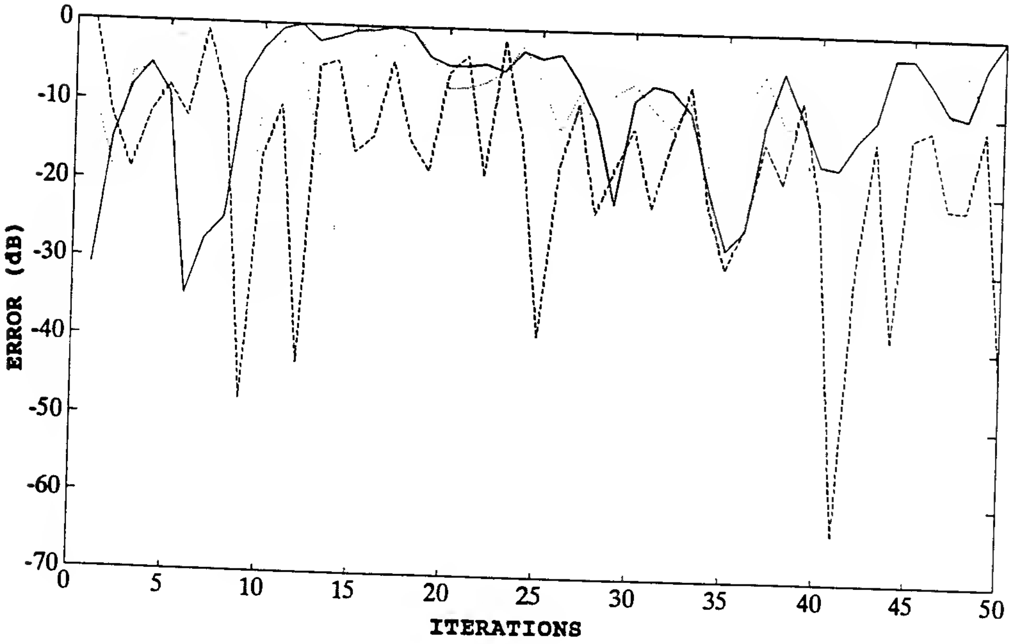


Figure 8: Close-up comparison between time domain LMS (solid), DCT LMS (dotted), and Wavelet domain LMS (dashed) assuming a DC10 aircraft RCS input.

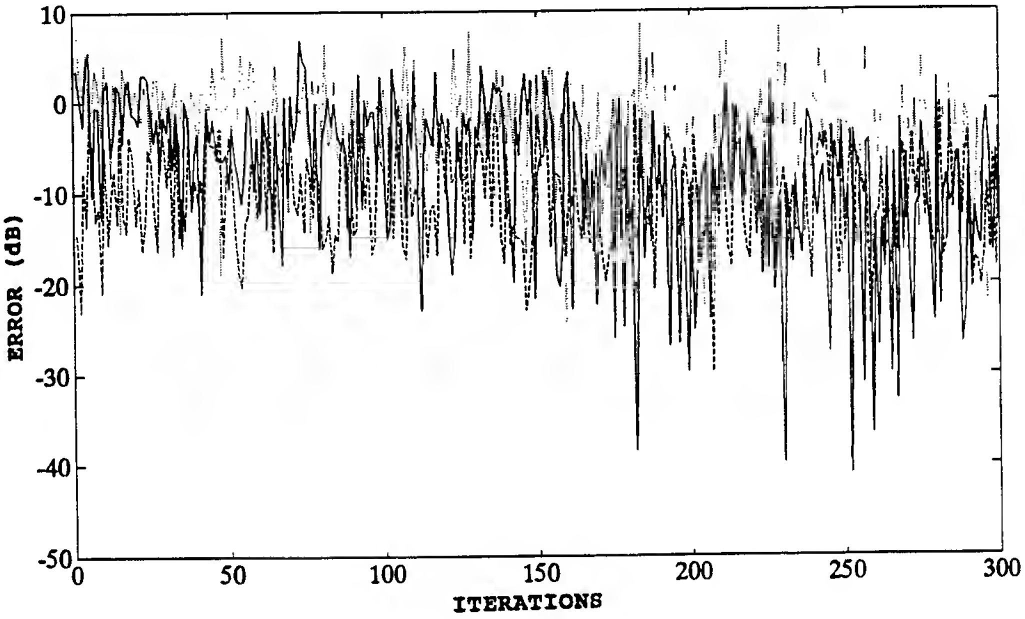


Figure 9: Comparison between time domain LMS (dotted), FFT LMS (solid), and wavelet domain (D19) LMS (dashed) assuming a sinusoidal input.

$$N_O = NK(3+\log_2 NK)+4 \tag{48}$$

$$N^{ct} = NK\log_2(NK)-1.5NK+4+(6NK+1) \tag{49}$$

$$N^{fft} = NK\log_2(NK)+(6NK+1) \tag{50}$$

where $N$, $N_O$, $N^{ct}$, $N^{fft}$ represent the number of multiplications needed for time domain LMS, self-orthogonalizing LMS, Cosine transform based LMS, and FFT based LMS. Accordingly the number of multiplications required by wavelet based LMS is

$$N^w = cNK+1 \tag{51}$$

where $6 \leq c \leq 7$. Clearly, the number of multiplications required by the wavelet transform LMS algorithm compares favorably with those of other transforms. The computational complexity of the above algorithms when the incoming signal is complex (with both quadrature and in-phase components which is the case in radar) is about seven to nine times that of real data but the computational burden of each of the above LMS techniques remains relatively the same.

## IV. Bandwidth Performance of Transform Domain Arrays

The bandwidth performance of an adaptive array algorithm is an important measure of its nulling power in the presence of wideband interference. Tapped delay line cancellers are usually introduced in an array to improve its bandwidth performance and studies suggest that the tapped delay line technique, though costly, does improve the nulling power of an array when wideband interference is present. Two studies [41] and [38] indicated that nulling using the tapped delay line approach is superior to employing an FFT prior to adaptation. In this section, we examine the significance of the wavelet subband decomposition in improving the nulling performance of an adaptive array when wideband interference is present. Our approach is similar to that of [41] which begins with a narrowband signal, interference, and noise model. The bandwidth of the interference signal is then increased and the nulling capability of the array is examined. Let the signal incident on the $k^{th}$ tap of the $m^{th}$ element of the array be

$$X_{mk} = X^d_{mk} + X^i_{mk} + X^n_{mk} \tag{52}$$

Note that, if $d(t)$ is the desired signal then

$$X^d_{mk} = d\left(t - [k-1]T_0 - [m-1]T_d\right) \tag{53}$$

with

$$T_d = \frac{L}{c}\sin\theta_d \tag{54}$$

where $L$ denotes the separation between array elements, and $c$ is the sped of light. The sampling period of the incoming signal is denoted by $T_0$. Similarly, the interference component is defined as

$$X^i_{mk} = i\left(t - [k-1]T_0 - [m-1]T_i\right) \tag{55}$$

where $i(t)$ is the interference signal arriving at an angle $\theta_i$, and

$$T_i = \frac{L}{c}\sin\theta_i \tag{56}$$

Let $p_d$ and $p_i$ denote the power of both the desired signal and the interference given by

$$p_d = E\{|d(t)|^2\} \tag{57}$$

$$p_i = E\{|i(t)|^2\} \tag{58}$$

similarly

$$X_{mk} = n_m(t - [k-1]T_0) \tag{59}$$

where $n_m$ is the noise component arriving at the $m^{th}$ array element. The covariance matrix $\Phi_X$ is then defined as $\Phi_d + \Phi_i + \Phi_n$

$$\begin{bmatrix} \Phi_{d_{11}} + \Phi_{i_{11}} + \Phi_{n_{11}} & \Phi_{d_{12}} + \Phi_{n_{12}} & \cdots & \Phi_{d_{1N}} + \Phi_{i_{1N}} \\ \Phi_{d_{21}} + \Phi_{i_{21}} & \Phi_{d_{22}} + \Phi_{i_{22}} + \Phi_{n_{22}} & \cdots & \Phi_{d_{2N}} + \Phi_{i_{2N}} \\ \cdots & \cdots & \cdots & \cdots \\ \Phi_{d_{N1}} + \Phi_{i_{N1}} & \cdots & \cdots & \Phi_{d_{NN}} + \Phi_{i_{NN}} + \Phi_{n_{NN}} \end{bmatrix} \tag{60}$$

where

$$\Phi_{d_{ml}} = E\{X_m^d X_l^{dT}\} \tag{61}$$

$$\Phi_{i_{ml}} = E\{X_m^i X_l^{iT}\} \tag{62}$$

$$\Phi_{n_{ml}} = E\{X_m^n X_l^{nT} \tag{63}$$

where $X_l^{dT}$ denotes the transpose of the desired signal vector received by the $l^{th}$ array element and so on. The output of the array $S$ is defined as [39]

$$s = W^T X = s_d + s_i + s_n \tag{64}$$

where

$$s_d = W^T X^d \tag{65}$$

$$s_i = W^T X^i \tag{66}$$

$$s_n = W^T X^n \tag{67}$$

where the vector $X^d$ is a cascade of the vectors $X_m^d$, $m = 1, \ldots, N$ and so on. The desired signal power is then

$$p_d = E\{|s(t)|^2\} = E\{W^T X^d X^{dT} W\} = W^T \Phi_d W \tag{68}$$

similarly

$$p_i = W^T \Phi_i W \tag{69}$$

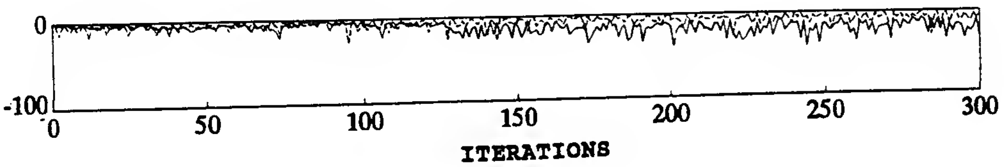


Figure 10: Comparison between time domain LMS (dotted) and wavelet domain LMS (solid), (the error averaged over 1000 experiments).

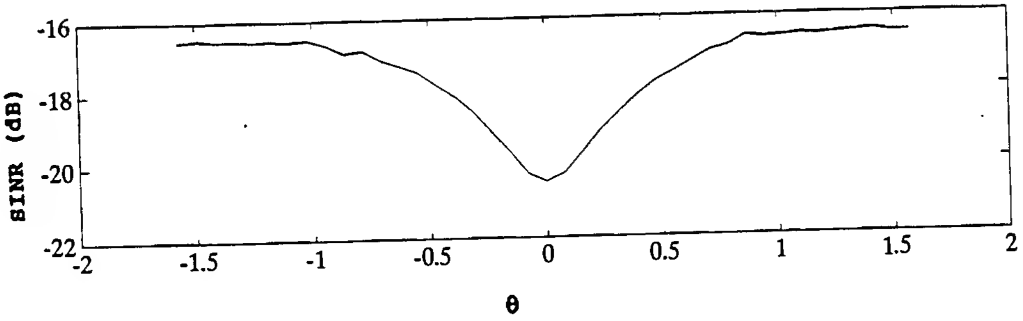


Figure 11: Bandwidth performance of an adaptive antenna array showing SINR vs interference angle $\theta_i$ $(-\frac{\pi}{2} \leq \theta_i \leq \frac{\pi}{2})$

$$p_n = W^T \Phi_n W \tag{70}$$

and the signal-to-(interference+noise) ratio at the output of the array is

$$SINR = \frac{p_d}{p_i + p_n} = \frac{W^T \Phi_d W}{W^T \left(\Phi_i + \Phi_n\right) W} \tag{71}$$

which can be computed by knowing $W$ and $s_d$, $s_i$ and $s_n$. The signals $d(t)$, $i(t)$ and $n(t)$, that are used in this study, were generated using autoregressive filtering of white Gaussian noise where the frequency response of the filter used is

$$H(\omega) = \Pi\left(\frac{\omega - \omega_0)}{\Delta\omega_0}\right) \tag{72}$$

where $\Pi(\omega/L)$ is a box function of base width $L$ and centered at $\omega = 0$. Therefore, the filtered signal represents a narrowband signal with relative bandwidth

$$B = \frac{\Delta\omega_0}{\omega_0} \tag{73}$$

The bandwidth of either the interference or the desired signal signal can be increased by increasing $\Delta\omega_0$. Notice that the impulse response of this filter is

$$h(k) = sinc\left(\frac{\Delta\omega_0 k}{2}\right) e^{j\omega_0 k}\Pi\left(\frac{k - M/2}{M}\right) \tag{74}$$

where $sinc(x) = sin(x)/x$. The impulse response is truncated to maintain causality. The nulling performance of the adaptive array is expected to deteriorate as the interference bandwidth is increased. The filtered signal generated using this algorithm can be presented directly to the array processor (as in Figure 4) or transformed into another domain using FFT, DCT, or wavelets and then processed as in Figure 5. Scenarios similar to those depicted in Figures 6, 7, 8, and 9. were attempted and the bandwidth performance of a two element array was examined. Arrays with 5, 6, and 12 elements were also examined and the output signal-to-(interference+noise) ratio SINR showed no evidence of any change in the bandwidth performance of an adaptive array despite the additional computational cost. Figure 10 shows the bandwidth performance (SINR versus the interference arrival angle $\theta_i$ which is directly related to frequency) of a two elements array with 64 tap delays. The incorporation of the wavelet transform, DCT, or the FFT did not improve the bandwidth performance of the array, and in fact the nulling capability of the array was slightly diminished upon transforming the input signal. This result agrees with Compton's work [39] which claims that no invertible transformation placed between the incoming signal and the array processor would improve the bandwidth performance of the array. Therefore, wavelet decomposition, being an invertible transformation, did not improve the nulling power of an array as the interference bandwidth was increased.

## V. Conclusions and Future Work

Faster adaptation rates can be achieved by inserting a wavelet transformer between the incoming signal and the LMS or Applebaum array processors. The transformation improves the condition number of the covariance matrix and thus improves the convergence rate of an array. The weights dynamic range, which is of practical interest, is also improved because of wavelet transformation by a factor directly proportional to the eigenvalues of the wavelet transformation matrix. The learning rate (or adaptation speed) of the LMS or Applebaum arrays can be further improved by using time dependent adaptation parameter $\mu$ or by using the self orthogonalization adaptation approach. Preliminary studies suggest that the weight convergence rate can be significantly improved using self orthogonalization with moderate increase in computational cost.

The incorporation of wavelet transformation in an adaptive antenna array does not enhance the array's capability of nulling wideband jamming beyond what can be achieved using tap delay line elements. This effect of wavelet transform on the nulling power of an array in the presence of wideband jamming requires further investigation assuming different signal plus interference scenarios and different wavelets.

The computational cost of employing wavelet transformation in adaptive arrays is modest and compares favorably with that of the FFT or the Cosine transform.

The success of the wavelet transform in improving the convergence rate or adaptation speed of an array is remarkable and deserves much further attention. An optimal wavelet suited for radar signals with wideband interference and multipath jamming is yet to be developed. A theoretical assessment of the underlying reasons for such an improvement in the convergence rate of adaptive arrays is yet to developed.

Acknowledgements: The author wishes to thank the Air Force Office of Naval Research for supporting this project. The author acknowledges and appreciates the support of fellow researchers at Wright Patterson Air Force Base, OH.

## References

[1] S. S. Narayan, A. M. peterson, and M. J. Narasimha, "Transform Domain LMS Algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 31, No. 3, pp 609-615, June 1983.

[2] N. erdol and F. Basbug, "Wavelet Transform Based Adaptive Filters: Analysis and New Results," submitted to *Applied and Computational Harmonic Analysis,* 1993.

[3] S. Hosur and A. H. tewfik, "Wavelet Transform Domain LMS Algorithm," *International Conference on Acoustics, Speech, and Signal Processing,* ICASSP '93, pp III-508-510, 1993.

[4] J. C. Lee and C. K. Un, "Performance of Transform Domain LMS Adaptive Digital Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 34, No.3, pp. 499-510, June 1986.

[5] C. Van den Branden Lamrecht and M. Karrakchou, "Subband Adaptive Filtering: The Mutual Wavelet Packets Approach," *International Conference on Acoustics, Speech, and Signal Processing,* ICASSP '93, pp. III-316 - III-319, 1993.

[6] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael eds., *Wavelets and their applications,* Jones and Bartlett, Boston, 1992.

[7] I.Jouny, "Target Description Using Wavelet Transform", *International Conference on Signal Processing,* San Francisco, CA, March 1992.

[8] S. G. Mallat, " A Theory for Multiresolution Signal Decomposition: The wavelet Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 11, No. 7, pp 674-693, July 1989.

[9] P. Flandrin, F. Magand, and M. Zakharia, "Generalized Target Description and Wavelet Decomposition," *IEEE Transactions on Signal Processing,* Vol. 38, No. 2, pp 350-352, February 1990.

[10] I. Daubechies, "Orthonormal Basis of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics,* Vol. XLI, pp 909-996, 1988.

[11] G. Strang, "Wavelets and Dilation Equations: A Brief Introduction," *SIAM,* Vol. 31, No. 4, pp 614-627, December 1989.

[12] O. Rioul and M. Vitterli, "Wavelets and Signal Processing," *IEEE Signal Processing Magazine,* pp 14-38, October 1991.

[13] R. Kronland-Martinet, J. Morlet, and A. Grossman, "Analysis of Sound Patterns Through Wavelet Transform," *International Journal of Pattern Recognition and Artificial Intelligence,* Vol. 1, No. 2, pp 273-302, 1987.

[14] Y. Meyer, "Orthonormal Wavelets," Wavelets, Time-Frequency Methods and Phase Space, *Proceedings of The International Conference*, Marsielle, France, Dec. 14-18, 1987, J.M. Combes et al. eds., Inverse Problems and Theoretical Imaging, Springer, pp 315, 1989.

[15] R. R. Coifman, "Wavelet Analysis and Signal Processing," in *Signal Processing, Part I: Signal Processing Theory*, L. Auslander et al. eds., IMA, Vol. 22, Springer, New York, 1990.

[16] O. Rioul and M. Vitterli, "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, pp 14-38, October, 1991.

[17] S. G. Mallat, "Multifrequency Channel Decompositions of Images and Wavelet Models," *IEEE Transactions on Signal Processing*, Vol. 37, No. 12, pp 2091-2110, December 1989.

[18] T. Edwards, "Discrete Wavelet Transform: Theory and Implementation," submitted to *IEEE Transactions on Signal Processing*, 1992.

[19] H. Resnikoff, "Wavelets and Adaptive Signal Processing," *Optical Engineering*, Vol. 31, No. 6, pp 1229-1234, June 1992.

[20] H. Szu, Yunlong Sheng, and Jing Chen, "Wavelet Transform as a Bank of Matched Filters," *Applied Optics*, Vol. 31, No. 17, pp 3267-3277, June 1992.

[21] Y. Zhang et al, "Optical Realization of Wavelet Transform for a One Dimensional Signal," *Optical Society of America*, Vol. 17, No. 3, pp 210-212, February 1992.

[22] M. C. Proenca, J. P. Rudant, and G. Flouzat, "Using Wavelets to Get SAR Images "FREE" of Speckle," *Proceedings IGARSS 92*, pp 887-889, Houston, May 1992.

[23] J. G. Teti and H. N. Kritikos, "Weyl-Heisenberg and Wavelet Coherent Frames for SAR Ocean Image Filtering," *Proceedings IGARSS 92*, pp 1318-1320, Houston, May 1992.

[24] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image Coding Using Wavelet Transform," *IEEE Transactions on Image Processing*, Vol. 1, No. 2, April 1992.

[25] S. Mallat and W. L. Hwang, "Singularity Detection and Processing with wavelets," *IEEE Transactions on Information Theory*, Vol. 38, No. 2, March 1992.

[26] S. Mallat and S. Zhong, "Characterization of Signals from Multiscale Edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 7, July 1992.

[27] S. Mallat, "Zero-Crossing and Wavelet Transform," *IEEE Transactions on Information Theory*, Vol. 37, No. 4, July 1991.

[28] H. Kim and H. Ling, "Analysis of Electromagnetic Backscattering Area Using Wavelets,"*Proceedings of IEEE AP/URSI Joint Symposium* , Vol. 4, pp 1877-1881, July 1992.

[29] I. Jouny, "Description and Recognition of Radar Targets Using Wavelets," *Final Report, AFOSR Summer Faculty Program*, August 1992.

[30] W. D. White, "Wideband Interference Cancellation in Adaptive Sidelobe Cancellers," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 19, No. 6, pp 915-924, November, 1993.

[31] F. W. Vook, R. T. Compton, "Bandwidth Performance of Linear Adaptive Arrays with Tapped Delay-Line Processing," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 28, No. 3, pp 901-908, July 1992.

[32] H. H. Szu, B. Telfer, and S. Kadambe, "Neural network adaptive wavelets for signal representation and classification," *Optical Engineering*, Vol 31, No. 9, pp 1907-1916, September, 1992.

[33] W. F. Gabriel, "Adaptive digital processing investigation of DFT subbanding vs transversal filter canceler," *Naval Research Laboratory technical report*, NRL report 8981, July 1986.

[34] W. F. Gabriel, "Adaptive Processing Array Systems," *Proceedings of The IEEE*, Vol. 80, No. 1, pp 152-162, January 1992.

[35] S. S. Narayan and A. M. Peterson, "Frequency Domain Least-Mean-Square Algorithm," *Proceedings of The IEEE*, Vol. 69, No. 1, pp 124-126, January 1981.

[36] J. T. Mayhan, A. J. Simmons, and W. C. Cummings, "Wide-Band Adaptive Nulling Using Tapped Delay Lines," *IEEE Transactions on Antennas and Propagation*, Vol. 29, No. 6, pp 923-936, November 1981.

[37] S. Mann and S. Haykin, "Adaptive Chirplet transform: an adaptive generalization of the wavelet transform," *Optical Engineering*, Vol. 31, No. 6, pp 1243-1256, June 1992.

[38] L. E. Brennan and I. S. Reed, "Adaptive Cancellation of Scattered Interference," *Adaptive Sensors, Inc.*, final report, December 1982.

[39] R. T. Compton, "The bandwidth performance of a two-element adaptive array tapped delay-line processing," *IEEE Transactions on Antennas and Propagation*, Vol. 36, No. 1, pp 5-13, January 1988.

[40] D. R. Morgan and A. Aridgides, "Adaptive Sidelobe Cancellation of Wide-Band Multipath Interference," *IEEE transactions on Antennas and Propagation*, Vol. 33, No. 8, pp 908-917, August 1985.

[41] R. T. Compton, "The Relationship Between Tapped Delay-Line and FFT Processing in Adaptive Arrays," *IEEE Transactions on Antennas and Propagation*, Vol. 36, No. 1, pp 15-26, January 1988.

# MICROMECHANICS OF MATRIX CRACKING IN BRITTLE MATRIX COMPOSITES

Autar K. Kaw
Associate Professor
Mechanical Engineering Department

University of South Florida
ENG 118, 4202 E. Fowler Avenue,
Tampa, FL 33620-5350

Final Report for:
Research Initiation Program
Wright-Patterson Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

University of South Florida

December 1993

# MICROMECHANICS OF MATRIX CRACKING IN BRITTLE MATRIX COMPOSITES

Autar K. Kaw
Associate Professor
Mechanical Engineering Department
University of South Florida

## Abstract

The effect of a frictional interface on the response of a unidirectional ceramic matrix composite under a remote axial tensile strain and a temperature change is studied. The geometry of the composite is approximated by a concentric cylinder model with an annular crack in the axial plane of the matrix. The fiber-matrix interface follows the Coulomb friction law. On applying the boundary and the interface continuity conditions, the solution is obtained in terms of coupled integral and linear equations, and inequality conditions.

The extent of the interfacial damage and the stress fields in the fiber and the matrix along the interface are studied for a SiC/CAS composite system for varying coefficient of friction, temperature change and remote uniform axial strain. These results are also compared with a shear lag analysis model for identical geometry and loading.

# MICROMECHANICS OF MATRIX CRACKING IN BRITTLE MATRIX COMPOSITES

Autar K. Kaw

## INTRODUCTION

Ceramic matrix composites are becoming attractive as load bearing structures for high temperature and corrosive atmosphere applications. Although these composites have higher ultimate strength and strain than monolithic ceramics, matrix cracking followed by interfacial failure is still a critical issue in their use.

Consider a unidirectional ceramic composite subjected to an axial strain along the fiber direction. The cracks will first develop in the matrix due to its lower failure strain than that of the fiber. When a matrix crack reaches the interface of the fiber and the matrix, the interface may open or slip. This opening/slipping of the interface blunts the crack, and slows and arrests the propagation of the crack. Although this blunting of the crack increases the fracture toughness of the composite, the damage in the interface reduces the axial compressive and transverse strength of the composite (Steif, 1984). Because of these conflicting effects of interfacial damage, it becomes important to fully understand the mechanics of matrix fracture in ceramic matrix composites as a function of material, geometrical and loading parameters.

Axisymmetric three dimensional failure mechanics models, which account for all equations of elasticity as well as assume an imperfect interface, for the fracture in ceramic matrix composites are reported in the literature. These include the work of Wijeyewickrema and Keer (1993), Kaw and Pagano (1993), and Schweitert and Steif (1991). The interface in all the above three studies is modeled differently.

Wijeyewickrema and Keer (1993) solved the problem of a composite cylinder made of a solid cylinder (fiber) bonded to a surrounding hollow cylinder (matrix) of finite outer radius. An annular crack was assumed in the matrix. The composite cylinder was subjected to a remote uniform tensile strain. The interface included a slip zone and was assumed to have a constant shear stress equal to the shear strength of the interface. This is a fairly valid assumption when the interfacial friction coefficient is small (Aksel, Hui and Lagoudas, 1991).

Kaw and Pagano (1993) solved for the same composite geometry as Wijeyewickrema and Keer (1993). Kaw and Pagano (1993) included an imperfect interface in the composite cylinder model but by approximating the interface by distributed shear springs of constant stiffness. Their model also included the effects of temperature change.

Schweitert and Steif (1991) used a similar geometry as the above two studies except two differences. First, the outer radius of the matrix was assumed to be infinite. Second, a penny shaped crack was assumed in the fiber (solid cylinder) instead of the annular crack in the matrix (hollow cylinder). They approximated the interface by the Coulomb friction law. The composite geometry was subjected to a pressure on the crack surface and a constant remote compressive radial stress. The pressure on the crack surface indirectly represented the matrix axial stresses due to a remote uniform axial strain. The remote radial stress represented residual stresses due to the mismatch of the linear coefficients of thermal expansion coefficient and the Poisson's ratio of the fiber and the matrix.

In the present study, several assumptions made in Schweitert and Steif's (1991) model are relaxed as follows.

- The dilute fiber volume fraction assumption is replaced by a nondilute fiber volume fraction.
- The fiber crack is replaced by an annular matrix crack. Also, the annular crack does not necessarily have to be a through crack. It can be internal, edge and/or touching the interface.
- The stresses due to the thermal expansion mismatch of the fiber and the matrix can be directly accounted in the model.

These relaxed assumptions allow direct study of the combined effect of material, thermomechanical loading and geometrical parameters. In the sections to follow is the formulation of the model. The effect of the coefficient of friction at the fiber-matrix interface, and the linear coefficients of thermal expansion of the fiber/matrix on the extent of interfacial damage, stress distribution at the interface, under a thermomechanical load are studied. These results are compared with an approximate model for an identical geometry and loading. The approximate model is based on axial stresses being independent of the radial coordinate similar to Gu and Mangonon's (1992) radially constrained matrix model.

## METHOD OF ANALYSIS

### Geometry

The geometry of the composite cylinder consists of an infinitely long fiber bonded to an annular matrix of finite outer radius (Figure 1). This geometry approximates a representative volume element (RVE) of a composite in a double hexagonal array. The cylindrical coordinates are denoted by r, θ and z, and $u_r$ and $u_z$ are the radial and axial displacements, respectively. The normal and shear stresses are denoted by $\sigma_{rr}$, $\sigma_{zz}$, $\sigma_{\theta\theta}$, and $\sigma_{rz}$. The indices 0 and 1 stand for the fiber and the matrix, respectively.

The fiber is approximated by a linearly elastic, isotropic, homogeneous and infinitely long solid cylinder of radius a, shear modulus $\mu_0$, Poisson's ratio $\upsilon_0$, Young's modulus $E_0=2(1+\upsilon_0)\mu_0$, and linear coefficient of thermal expansion $\alpha_0$. The matrix is approximated by a linearly elastic, isotropic, homogeneous and infinitely long annular cylinder of inner radius a and outer radius c, shear modulus $\mu_1$, Poisson's ratio $\upsilon_1$, Young's modulus, $E_1 = 2(1 + \upsilon_1)\mu_1$ and linear coefficient of thermal expansion, $\alpha_1$. An annular crack of length 'e-d' ($a \le d < e \le c$) in the z=0 plane, at a distance of 'd-a' from the interface is assumed in the matrix. The fiber volume fraction is $V_f=a^2/c^2$.

### Boundary and Continuity Conditions

The composite cylinder is subjected to a monotonically increasing axial remote strain, $\epsilon_0$ on the ends plus a constant temperature change, ΔT.

The imperfect interface between the fiber and the matrix follows the Coulomb friction law and may have open, slip and stick zones.

The length of the open zone is '$z_1$', while the length of the slip zone is '$z_2$-$z_1$'. The kinetic and static friction coefficients are considered to be equal. The friction coefficient '$\rho$' is assumed to be constant in the slip zone. The superscripts '0' and '1' denote the fiber and the matrix, respectively. The continuity conditions at the interface between the fiber and the matrix at r=a are, hence, given by

$$\sigma^0_{rr}(a,z)=\sigma^1_{rr}(a,z)\,,\ 0\le|z|<\infty, \tag{1.a}$$

$$\sigma^0_{rz}(a,z)=\sigma^1_{rz}(a,z)\,,\ 0\le|z|<\infty. \tag{1.b}$$

Also, at the interface (r=a) between the fiber and the matrix, the zones are governed by

Open zone

The crack surfaces are traction free as given by

$$\sigma^0_{rz}(\mathrm{a},z)=0,\quad 0\le|z|\le z_1, \tag{2.a}$$

$$\sigma^0_{rr}(\mathrm{a},z)=0,\quad 0\le|z|\le z_1, \tag{2.b}$$

constrained by the crack is open

$$u^1_r(\mathrm{a},z)-u^0_r(\mathrm{a},z)>0,\quad 0\le|z|<z_1, \tag{2.c}$$

Slip zone

Radial contact is maintained

$$u^0_r(\mathrm{a},z)=u^1_r(\mathrm{a},z),\ z_1\le|z|\le z_2, \tag{2.d}$$

and the shear stress is related to the radial stress through the coefficient of friction '$\rho$'.

$$\sigma^0_{rz}(\mathrm{a},z)=-\rho\ \sigma^0_{rr}(\mathrm{a},z),\ z_1\le|z|\le z_2, \tag{2.e}$$

There needs to be a positive dissipation of energy in the slip zone implying the direction of the shear stress and increment in axial slip as

$$sgn(\frac{d}{dt}[u^1_z(\mathrm{a},z)-u^0_z(\mathrm{a},z)])=sgn[\sigma^0_{rz}(\mathrm{a},z)],\ z_1\le|z|\le z_2, \tag{2.f}$$

The variable 't' is a time-like parameter and is assumed to increase monotonically with increasing remote axial strain, $\epsilon_0$. The constraining conditions include that the radial stress is compressive

$$\sigma^0_{rr}(\mathrm{a},z)<0,\quad z_1<|z|<z_2. \tag{2.g}$$

Stick zone

The radial and axial displacements are continuous at the interface

$$u^0_r(\mathrm{a},z)=u^1_r(\mathrm{a},z),\quad z_2\le|z|<\infty, \tag{2.h}$$

$$u^0_z(\mathrm{a},z)=u^1_z(\mathrm{a},z),\ z_2\le|z|<\infty, \tag{2.i}$$

constrained by the radial stress is compressive

$$\sigma^0_{rr}(\mathrm{a},z)<0,\quad z_2\le|z|<\infty, \tag{2.j}$$

and the absolute value of the shear stress is such that it does not allow slip as

$$|\sigma^0_{rz}(\mathrm{a},z)|<-\rho\ \sigma^0_{rr}(\mathrm{a},z),\quad z_2<|z|<\infty, \tag{2.k}$$

The boundary conditions at the matrix edge r=c are given by

$$u_r^1(c,z) = u_r^{1T}(c,z) + u_r^{1e}(c,z),\ 0\leq|z|<\infty, \tag{3}$$

$$\sigma_{rz}^1(c,z) = 0,\ 0\leq|z|\infty, \tag{4}$$

where $u_r^{1T}$ is the radial displacement in the uncracked composite due to a temperature change, ΔT; $u_r^{1e}$ is the radial displacement in the uncracked composite due to a remote axial strain, $\epsilon_0$ (Kaw and Pagano, 1993; see their Appendix A and B).

The boundary condition (3) results in the slope of the crack surface $\partial u_z^1(r,0)/\partial r|_{r=e}$ equal to zero for the edge crack problem (e=c). Also at the edge r=c, far away from the crack plane, the radial stress becomes zero.

The shear stress in the composite cylinder at the crack plane z=0 is

$$\sigma_{rz}^0(r,0) = 0, r\leq a, \tag{5.a}$$

$$\sigma_{rz}^1(r,0) = 0, a\leq r\leq c. \tag{5.b}$$

The other boundary conditions at the crack plane z=0 are

$$u_z^0(r,0) = 0, r\leq a, \tag{6.a}$$

$$u_z^1(r,0) = 0, a\leq r<d,\ e<r\leq c, \tag{6.b}$$

$$\sigma_{zz}^1(r,0) = 0, d<r<e, \tag{6.c}$$

constrained by the transverse crack is open as

$$u_z^1(r,0) > 0,\ d<r<e. \tag{7}$$

## Formulation

The solution for the above problem is obtained by a related boundary element method. The composite cylinder geometry of Figure 1 can be viewed as two auxiliary bodies as shown in Figure 2.

The first free body diagram is that of a solid cylinder with unknown surface tractions on the boundary r=a. The second free body diagram is that of a hollow cylinder with unknown surface tractions on the inner radius r=b, (b=a in the composite cylinder) unknown slope of the crack opening displacement at z=0, and known boundary conditions on the outer radius r=c.

The complete displacement and stress fields of the solid cylinder (fiber) can be now found in terms of the unknown surface tractions at r=a. For the hollow cylinder (matrix), the complete displacement and stress fields can also be found in terms of the unknown tractions at r=b and the unknown slope of the crack opening displacement at z=0. Once these field equations are obtained, the continuity and boundary conditions (1-7) can be applied to find the solution in terms of coupled integral and linear equations. These coupled equations can then be solved numerically to find the stress/displacement field in the entire composite geometry.

Field equations for the solid cylinder

The displacement/stress field for an axisymmetric solid cylinder of radius 'a', shear modulus $\mu_0$, Poisson's ratio $\nu_0$ and which has symmetry about the z=0 plane with boundary stresses

$$\sigma_{rr}^0(a,z) = S_0(z), 0<z<\infty, \tag{8.a}$$

$$\sigma_{rz}^0(a,z) = T_0(z), 0<z<\infty, \tag{8.b}$$

where $S_0(z)$ and $T_0(z)$ are absolutely integrable in $(0,\infty)$, is given by

$$\begin{aligned} M_i^0(r,z) = {} & \frac{2}{\pi}\int_0^\infty [k_{1i}(r,s) - k_{1i}^\infty(r,s)]\, Cossin(zs)\, ds \int_0^\infty T_0(t)\, Sin(st)\, dt \\ & + \frac{2}{\pi}\int_0^\infty T_0(t)\, dt \int_0^\infty k_{1i}^\infty(r,s)\, Cossin(zs)\, Sin(st)\, ds \\ & + \frac{2}{\pi}\int_0^\infty [k_{2i}(r,s) - k_{2i}^\infty(r,s)]\, Cossin(zs)\, ds \int_0^\infty S_0(t)\, Cos(st)\, dt \\ & + \frac{2}{\pi}\int_0^\infty S_0(t)\, dt \int_0^\infty k_{2i}^\infty(r,s)\, Cossin(zs)\, Cos(st)\, ds, \quad i=1,\dots,6, \end{aligned} \tag{9}$$

where

$M_1$ = Radial displacement, $u_r(r,z)$,

$M_2$ = Axial displacement, $u_z(r,z)$,

$M_3$ = Radial stress, $\sigma_{rr}(r,z)$,

$M_4$ = Axial stress, $\sigma_{zz}(r,z)$,

$M_5$ = Shear stress, $\sigma_{rz}(r,z)$,

$M_6$ = Hoop stress, $\sigma_{\theta\theta}(r,z)$,

and Cossin (zs) = Cos (zs), if i = 1, 3, 4, 6,
= Sin (zs), if i = 2, 5.

The expressions for $k_{1i}$ and $k_{2i}$ are given by

$$k_{\ell i} = \sum_{j=1}^{2} E_{ij}(r,s)\ f_{j\ell}(s),\ \ell=1,2 \text{ and } i=1,2,3,4,5, \tag{10}$$

$$f_{j\ell} = [\alpha]^{-1}_{2x2}\ [\gamma]_{2x2}\ ,j=1,2 \text{ and } \ell=1,2. \tag{11}$$

$$\alpha_{11} = \bar{I}_1(as)$$

$$\alpha_{12} = as\ \bar{I}_0(as) + 2(1-\nu_0)\ \bar{I}_1(as) \tag{12}$$

$$\alpha_{21} = -\bar{I}_0(as) + \bar{I}_1(as)/(as)$$

$$\alpha_{22} = (2\nu_0-1)\bar{I}_0(as) - (as)\bar{I}_1(as)$$

$$\gamma = \frac{1}{s^3}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},$$

$$E_{11} = -\frac{1}{2\mu_0}s^2\ \bar{I}_1(rs)\,e^{-(a-r)s}, E_{12} = -\frac{1}{2\mu_0}s^2\ rs\ \bar{I}_0(rs)\,e^{-(a-r)s},$$

$$E_{21} = \frac{1}{2\mu_0}s^2\bar{I}_0(rs)\,e^{-(a-r)s},\ E_{22} = \frac{1}{2\mu_0}s^2[4(1-\nu_0)\bar{I}_0(rs)+rs\bar{I}_1(rs)]\,e^{-(a-r)s},$$

$$E_{31} = s^3[-\bar{I}_0(rs)+\bar{I}_1(rs)/rs]\,e^{-(a-r)s},\ E_{32} = s^3[(2\nu_0-1)\bar{I}_0(rs) - rs\bar{I}_1(rs)]\,e^{-(a-r)s},$$

$$E_{41} = s^3\ \bar{I}_0(rs)\,e^{-(a-r)s},\ E_{42} = s^3[2(2-\nu_0)\bar{I}_0(rs) + rs\bar{I}_1(rs)]\,e^{-(a-r)s}, \tag{13}$$

$$E_{51} = s^3\bar{I}_1(rs)\,e^{-(a-r)s},\ E_{52} = s^3[rs\bar{I}_0(rs) + 2(1-\nu_0)\bar{I}_1(rs)]\,e^{-(a-r)s},$$

$$\bar{I}_i(x) = e^{-x}I_i(x). \tag{14}$$

where $I_i(x)$, i=0,1 is the hyperbolic Bessel's function.

The above expressions have been obtained by solving the boundary value problem with boundary conditions given by using the Love's stress functions (Wijeyewickrema and Keer, 1991) defined for the axisymmetric torsionless infinite solid cylinder. The reason for writing the expressions in the above form (Equation (9)) is that one can reduce the double integrals to single integrals with a judicious choice of the unknown traction functions $S_0$ and $T_0$, and also by reducing the $k_{\ell i}^{\infty}$ as a product of $e^{-(a-r)s}$ and simple polynomials of (s) (some $k_{\ell i}^{\infty}$ for displacements also require taking 1/s terms in addition to the polynomial terms for faster convergence).

The $k_{\ell i}^{\infty}$ terms are evaluated by finding $k_{\ell i}^{\infty}$ in closed form as $s\to\infty$ by using the asymptotic expansion for $I_i(x)$ for large values of 'x' as (Abramowitz and Stegun, 1970; Page 377)

$$I_i(x) = \left(1 - \frac{\delta-1}{8x} + \frac{(\delta-1)(\delta-9)}{2(8x)^2} - \ldots\right)\frac{e^x}{\sqrt{2\pi x}}, i = 0,1,2, \delta=4i^2. \qquad (15)$$

**Field equations for the hollow cylinder**

The displacement/stress field for an axisymmetric hollow cylinder of inner radius 'b', outer radius 'c', shear modulus $\mu_1$, Poisson's ratio $\upsilon_1$, symmetric about the z=0 plane with boundary conditions

$$\sigma_{rr}^1\ (b,z) = S_1(z),\ 0\le z<\infty, \qquad (16.a)$$

$$\sigma_{rz}^1(b,z) = T_1(z),\ 0\le z<\infty, \qquad (16.b)$$

$$u_r^1(c,z) = 0,\ 0\le z<\infty, \qquad (16.c)$$

$$\sigma_{rz}^1(c,z) = 0,\ 0\le z<\infty, \qquad (16.d)$$

$$\frac{\mu_1}{1-\nu_1}\frac{\partial}{\partial r}u_z^1(r,0) = \phi(r),\ d<r<e, \qquad (16.e)$$

$$u_z^1(r,0) = 0,\ b<r<d,\ e<r<c, \qquad (16.f)$$

where $S_1(z)$ and $T_1(z)$ are absolutely integrable in $(0,\infty)$ and $\phi(r)$ is integrable in (d,e) are given by

$$
\begin{aligned}
M_i^1(r,z) = {} & \int_0^{\infty}[p_{1i}(r,s) - p_{1i}^{\infty}(r,s)]\, Cossin(zs)\, ds \int_0^{\infty} T_1(t)\, Sin(st)\, dt \\
& + \int_0^{\infty} T_1(t)\, dt \int_0^{\infty} p_{1i}^{\infty}(r,s)\, Cossin(zs)\, Sin(st)\, ds \\
& + \int_0^{\infty} S_1(t)\, dt \int_0^{\infty} p_{2i}^{\infty}(r,s)\, Cossin(zs)\, Cos(st)\, ds \qquad (17) \\
& + \int_0^{\infty}[p_{2i}(r,s) - p_{2i}^{\infty}(r,s)]\, Cossin(zs)\, ds \int_0^{\infty} S_1(t)\, Cos(st)\, dt \\
& + \int_d^{e} \phi(t)\, dt \int_0^{\infty}[p_{3i}(r,s,t) - p_{3i}^{b\infty}(r,s,t) \\
& \qquad - p_{3i}^{c\infty}(r,s,t)]\, Cossin(zs)\, ds \\
& + \int_d^{e} \phi(t)\, dt \int_0^{\infty}\{[p_{3i}^{b\infty}(r,s,t) + p_{3i}^{c\infty}(r,s,t)]\, Cossin(zs)\, ds \\
& \qquad + N_i(r,t,z)\}, \quad i=1,\ldots,6
\end{aligned}
$$

where,

$$p_{\ell i} = \sum_{j=1}^{4} F_{ij}(r,s)\; g_{j\ell}(s), \quad \ell=1,2,3 \text{ and } i=1,2,3,4,5,6, \qquad (18)$$

$$g_{j\ell} = [\beta]_{4x4}^{-1}\; [\Gamma]_{4x3}\;, j=1,2,3,4 \text{ and } \ell=1,2,3. \qquad (11)$$

$$\Gamma = \frac{1}{s^3}\begin{bmatrix} 1 & 0 & h_1(s,t) \\ 0 & 1 & h_2(s,t) \\ 0 & 0 & h_3(s,t) \\ 0 & 0 & h_4(s,t) \end{bmatrix}, \qquad (20)$$

$$\beta_{11} = \overline{I_1}(bs)\; e^{-(c-b)s}, \qquad (21.a)$$

$$\beta_{12} = \left[bs\,\overline{I_0}(bs) + 2(1-\nu_1)\,\overline{I_1}(bs)\right] e^{-(c-b)s}, \qquad (21.b)$$

$$\beta_{13} = -\overline{K_1}(bs)\,, \qquad (21.c)$$

$$\beta_{14} = -bs\,\overline{K_0}(bs) + 2(1-\nu_1)\,\overline{K_1}(bs)\,, \qquad (21.d)$$

$$\beta_{21} = \left[-\overline{I_0}(bs) + \overline{I_1}(bs)/(bs)\right] e^{-(c-b)s}, \tag{21.e}$$

$$\beta_{22} = -\left[(1 - 2\nu_1)\,\overline{I_0}(bs) + bs\,\overline{I_1}(bs)\right] e^{-(c-b)s}, \tag{21.f}$$

$$\beta_{23} = -\left[\overline{K_0}(bs) + \overline{K_1}(bs)/(bs)\right], \tag{21.g}$$

$$\beta_{24} = (1 - 2\nu_1)\,\overline{K_0}(bs) - bs\,\overline{K_1}(bs), \tag{21.h}$$

$$\beta_{31} = -\overline{I_1}(cs), \tag{21.i}$$

$$\beta_{32} = -cs\,\overline{I_0}(cs), \tag{21.j}$$

$$\beta_{33} = \overline{K_1}(cs)\; e^{-(c-b)s}, \tag{21.k}$$

$$\beta_{34} = cs\,\overline{K_0}(cs)\; e^{-(c-b)s}, \tag{21.l}$$

$$\beta_{41} = \overline{I_1}(cs), \tag{21.m}$$

$$\beta_{42} = cs\,\overline{I_0}(cs) + 2(1 - \nu_1)\,\overline{I_1}(cs), \tag{21.n}$$

$$\beta_{43} = -\overline{K_1}(cs)\; e^{-(c-b)s}, \tag{21.o}$$

$$\beta_{44} = \left[-cs\,\overline{K_0}(cs) + 2(1 - \nu_1)\,\overline{K_1}(cs)\right] e^{-(c-b)s}, \tag{21.p}$$

$$F_{11} = -\frac{1}{2}\mu_1 \overline{I_1}(rs)\, e^{-s(c-r)}, \tag{22.a}$$

$$F_{12} = rs\,\overline{I_0}(rs)\, e^{-s(c-r)}, \tag{22.b}$$

$$F_{13} = \overline{K_1}(rs)\, e^{s(b-r)}, \tag{22.c}$$

$$F_{14} = rs\,\overline{K_0}(rs)\, s^2 e^{s(b-r)}, \tag{22.d}$$

$$F_{21} = \frac{1}{2}\mu_1 \overline{I_0}(rs)\, e^{-s(c-r)}, \tag{22.e}$$

$$F_{22} = [4(1 - \nu_1)\overline{I_0}(rs) + rs\overline{I_1}(rs)]\, e^{-s(c - r)}, \tag{22.f}$$

$$F_{23} = \overline{K_0}(rs)\, e^{s(b - r)}, \tag{22.g}$$

$$F_{24} = [rs\overline{K_1}(rs) - 4(1 - \nu_1)\overline{K_0}(rs)]\, e^{s(b - r)}, \tag{22.h}$$

$$F_{31} = [-\overline{I_0}(rs) + \overline{I_1}(rs)/rs]\, e^{-s(c - r)}, \tag{22.i}$$

$$F_{32} = [(2\mu_0 - 1)\overline{I_0}(rs) - rs\overline{I_1}(rs)]\, e^{-s(c - r)}, \tag{22.j}$$

$$F_{33} = [\overline{K_0}(rs)\ \overline{K_1}(rs)/(rs)]\, e^{s(b - r)}, \tag{22.k}$$

$$F_{34} = [(1 - 2\nu_1)\overline{K_0}(rs) - rs\overline{K_1}(rs)]\, e^{s(b - r)}, \tag{22.l}$$

$$F_{41} = \overline{I_0}(rs)\, e^{-s(c - r)}, \tag{22.m}$$

$$F_{42} = [2(2-\nu_1)\overline{I_0}(rs) + rs\overline{I_1}(rs)]\, e^{-s(c - r)}, \tag{22.n}$$

$$F_{43} = \overline{K_0}(rs)\, e^{s(b - r)}, \tag{22.o}$$

$$F_{44} = [-2(2-\nu_1)\overline{K_0}(rs) + rs\overline{K_1}(rs)]\, e^{s(b - r)}, \tag{22.p}$$

$$F_{51} = \overline{I_1}(rs)\, e^{-s(c - r)}, \tag{22.q}$$

$$F_{52} = [rs\overline{I_0}(rs) + 2(1 - \nu_1)\overline{I_1}(rs)]\, e^{-s(c - r)}, \tag{22.r}$$

$$F_{53} = \overline{K_1}(rs)\, e^{s(b - r)}, \tag{22.s}$$

$$F_{54} = [2(1-\mu_1)\overline{K_1}(rs) - rs\overline{K_0}(rs)]\, e^{s(b - r)}, \tag{22.t}$$

$$F_{6j} = \frac{2\mu_1(1 - \nu_1)}{r} F_{1j} + \nu_1(F_{3j} + F_{4j}), \quad j=1,2,3,4, \tag{22.u}$$

$$\overline{K_i}(x) = e^{x} K_i(x), \tag{23}$$

where $K_i(x)$, i=0,1 are the Bessel's functions of second kind.

$$\phi(r) = \frac{\mu_1}{1-\nu_1}\frac{\partial}{\partial r}M_2(r,0),\ d<r<e, \tag{24}$$

$$h_1(t,s)=-s\left[bsI_0(bs)K_1(ts)-tsI_1(bs)K_0(ts)\right]e^{-(t-b)s}, \tag{25.a}$$

$$\begin{aligned} h_2(t,s)=-s[tsI_0(bs)K_0(ts)+I_0(bs)K_1(ts)\ -\ \frac{t}{b}I_1(bs)K_0(ts) \\ -[bs+\frac{2(1-\nu_1)}{bs}]I_1(bs)K_1(ts)]\ e^{-(t-b)s}, \end{aligned} \tag{25.b}$$

$$\begin{aligned} h_3(t,s)=-s[-tsI_0(ts)K_1(cs)+csI_1(ts)K_0(cs)+2(1-\nu_1)I_1(ts) \\ K_1(cs)]\,e^{-(c-t)s}, \end{aligned} \tag{25.c}$$

$$h_4(t,s)=-s\left[-tsI_0(ts)K_1(cs)-csI_1(ts)K_0(cs)\right]e^{-(c-t)s}, \tag{25.d}$$

$$N_1(r,t,z)\ =\ t\ [(1-2\nu_1)I_{01}-zI_{11}], \tag{26.a}$$

$$N_2(r,t,z)\ =\ t\ [2(1-\nu_1)I_{00}+zI_{10}]\ -\ \frac{(1-\nu_1)}{\mu_1}, \tag{26.b}$$

$$N_3(r,t,z)\ =\ t\ [I_{01}-zI_{02}\ -\frac{(1-2\nu_1)}{r}\ I_{10}\ -\frac{z}{r}\ I_{11}], \tag{26.c}$$

$$\begin{aligned} N_4(r,t,z)\ &=\ t\ [I_{01}\ +\ z\ I_{02}],z>0, \\ &=\frac{m(r,t)-1}{t-r}\ +\ \frac{m(r,t)}{t+r},\ z=0, \end{aligned} \tag{26.d}$$

$$\begin{aligned} m(r,t)\ &=\ E(\frac{r}{t}),\ r<t, \\ &=\frac{r}{t}E(\frac{t}{r})\ +\ \frac{t^2\ -\ r^2}{rt}K(\frac{t}{r}),r>t, \end{aligned} \tag{26.e}$$

$$N_5(r,t,z)\ =\ tzI_{12}, \tag{26.f}$$

$$N_6(r,t,z)\ =\ t\ [\frac{(1-2\nu_1)}{r}I_{10}\ -\ \frac{z}{r}\ I_{11}\ +\ 2\nu_1], \tag{26.g}$$

$$I_{ij} = \int_0^{\infty} p^j e^{-zp} J_1(pt)\; J_i(rp)\; dp. \qquad (27)$$

Equation (17) has been obtained by using the expression given for stresses and displacements in an infinite isotropic hollow cylinder (Erdol and Erdogan, 1978) and applying boundary conditions given by (16). The integrals in equation (27) can be simplified in terms of the elliptic integrals of first, second and third kind. The $p_{\ell i}^{\infty}$ are evaluated by finding $p_{\ell i}$ as $s\to\infty$ by using the asymptotic expansion for $I_i$ (x) as given by equation (18) and $K_i$ (x) for large values of x as (Abramowitz and Stegun, 1970, page 377).

$$K_i(x) = \left(1 + \frac{\delta - 1}{8x} + \frac{(\delta-1)(\delta-9)}{2(8x)^2} + \;.\;.\;.\right)\frac{e^{-x}}{\sqrt{2\pi x}}, \quad i=0,1,\delta = 4i^2. \qquad (28)$$

The $p_{3i}^{b\infty}$ and $p_{3i}^{c\infty}$ are asymptotic values of $p_{3i}(r, s\to\infty)$ as $r\to b$ and $r\to c$, respectively.

The above stress and displacement field equations (9) and (17) for the solid and hollow cylinder are restricted by the condition of absolute integrability of the tractions. These field equations cannot hence directly represent non-vanishing stresses due to temperature change and remote uniform strain as $z\to\infty$. However, the stresses as $z\to\infty$ are independent of the matrix and interfacial damage, and are the stresses and displacements due to the temperature change and remote strain in the undamaged composite cylinder. Hence, the complete stress and displacement field of the solid and the hollow cylinder in the presence of a temperature change and remote uniform strain in the composite cylinder is given by

$$u_r^j(r,z) = M_1^j(r,z) + u_r^{jT}(r,z) + u_r^{j\epsilon}(r,z), \quad j=0,1, \qquad (29.a)$$

$$u_z^j(r,z) = M_2^j(r,z) + u_z^{jT}(r,z) + u_z^{j\epsilon}(r,z), \quad j=0,1, \qquad (29.b)$$

$$\sigma_{rr}^j(r,z) = M_3^j(r,z) + \sigma_{rr}^{jT}(r,z) + \sigma_{rr}^{j\epsilon}(r,z), \quad j=0,1, \qquad (29.c)$$

$$\sigma_{zz}^j(r,z) = M_4^j(r,z) + \sigma_{zz}^{jT}(r,z) + \sigma_{zz}^{j\epsilon}(r,z), \quad j=0,1, \qquad (29.d)$$

$$\sigma^j_{rz}(r,z) = M^j_5(r,z), \quad j=0,1, \tag{29.e}$$

$$\sigma^j_{\theta\theta}(r,z) = M^j_6(r,z) + \sigma^{jT}_{\theta\theta}(r,z) + \sigma^{j\epsilon}_{\theta\theta}(r,z), \quad j=0,1. \tag{29.f}$$

In the above equations (29), the second and third terms on the right hand side are the known stresses and displacements due to the temperature change, ΔT and axial remote strain, $\epsilon_0$ respectively, in the undamaged composite cylinder and are given in Kaw and Pagano (1993) (see their Appendices A and B). The superscript 'T' corresponds to the effect of temperature change, $\Delta T$; the superscript '$\epsilon$' corresponds to the effect of the axial remote tensile strain, $\epsilon_0$.

Now the interface and boundary conditions (1-7) can be applied to give integral equations with $S_0$, $T_0$, $S_1$ and $T_1$, $\phi$ as the five unknown functions. These equations can be solved simultaneously to find the five unknown functions. The numerical scheme to find these functions is given in the next section.

## NUMERICAL SCHEME

The numerical scheme is discussed first for the case of an internal edge crack not touching the interface (d>a, e=c). The changes required for the numerical scheme for other cases follow at the end of this section.

### Internal edge crack (d>a, e=c)

The application of the boundary and continuity conditions as given by equations (1-7) should give the values of the interface stresses, $T_0$, $S_0$, $T_1$, and $S_1$ and the slope of the crack opening displacement function, $\phi$. This is done as follows.

Assume that the range of $0<z<\infty$ is divided into n unequal segments

$$\begin{aligned} T_0(z) &= A_i + B_i z, \quad \omega_i<z<\omega_{i+1}, \quad i=1,2\ .\ .\ .\ .n-1, \\ &= B_n/z^3, \quad \omega_n<z<\infty, \end{aligned} \tag{30.a}$$

$$\begin{aligned} S_0(z) &= C_i + D_i z, \quad \omega_i<z<\omega_{i+1}, \quad i=1,2\ .\ .\ .\ .n-1, \\ &= D_n/z^3, \quad \omega_n<z<\infty. \end{aligned} \tag{30.b}$$

$$\begin{aligned} T_1(z) &= P_i + Q_i z, \quad \omega_i<z<\omega_{i+1}, \quad i=1,2\ .\ .\ .\ .n-1, \\ &= Q_n/z^3, \quad \omega_n<z<\infty, \end{aligned} \tag{30.c}$$

$$S_1(z) = R_i + S_i z, \quad \omega_i<z<\omega_{i+1}, \quad i=1,2 \ldots .n-1,$$
$$= S_n/z^3, \quad \omega_n<z<\infty. \tag{30.d}$$

Further assume that the range (d,e) of the crack is divided into m unequal segments such that $r_1$=d and $r_{m+1}$=e, and

$$\phi(r) = (U_i+V_i r)\,w(r), \quad r_i<z<r_{i+1}, \quad i=1,2.....,m, \tag{31.a}$$

where

$$w(r) = \frac{1}{\sqrt{r-d}}, \tag{31.b}$$

is the weight function denoting the singularity of the slope of the crack opening displacement at the crack tip (r=d). The problem, hence, reduces to finding the values of the constants $A_i$, $B_i$, $C_i$, $D_i$, $P_i$, $Q_i$, $R_i$, $S_i$, $U_i$, and $V_i$ in equations (30) and (31).

Substituting the expressions for the tractions and slope given by equations (30) and (31) in equations (9) and (17), the radial and axial displacement at the interface (r=a) in the two bodies 0 and 1, and the axial stress in body 1 on the z=0 plane can be written as

$$u_r^0(a,z) = \sum_{i=1}^{n-1} A_i X_{i1}(z) + \sum_{i=1}^{n} B_i X_{i2}(z) + \sum_{i=1}^{n-1} C_i X_{i3}(z) + \sum_{i=1}^{n} D_i X_{i4}(z), \tag{32.a}$$

$$u_z^0(a,z) = \sum_{i=1}^{n-1} A_i X_{i5}(z) + \sum_{i=1}^{n} B_i X_{i6}(z) + \sum_{i=1}^{n-1} C_i X_{i7}(z) + \sum_{i=1}^{n} D_i X_{i8}(z), \tag{32.b}$$

$$u_r^1(b,z) = \sum_{i=1}^{n-1} P_i Y_{i1}(z) + \sum_{i=1}^{n} Q_i Y_{i2}(z) + \sum_{i=1}^{n-1} R_i Y_{i3}(z) + \sum_{i=1}^{n} S_i Y_{i4}(z)$$
$$+ \sum_{i=1}^{m} U_i Y_{i5}(z) + \sum_{i=1}^{m} V_i Y_{i6}(z), \tag{32.c}$$

$$u_z^1(b,z) = \sum_{i=1}^{n-1} P_i Y_{i7}(z) + \sum_{i=1}^{n} Q_i Y_{i8}(z) + \sum_{i=1}^{n-1} R_i Y_{i9}(z) + \sum_{i=1}^{n} S_i Y_{i10}(z)$$
$$+ \sum_{i=1}^{m} U_i Y_{i11}(z) + \sum_{i=1}^{m} V_i Y_{i12}(z), \tag{32.d}$$

$$\sigma^1_{zz}(r,0) = \sum_{i=1}^{n-1} P_i Z_{i1}(r) + \sum_{i=1}^{n} Q_i Z_{i2}(r) + \sum_{i=1}^{n-1} R_i Z_{i3}(r) + \sum_{i=1}^{n} S_i Z_{i4}(r) + \sum_{i=1}^{m} U_i Z_{i5}(r) + \sum_{i=1}^{m} V_i Z_{i6}(r), \tag{32.e}$$

where $X_{ij}$ and $Y_{ij}$ are functions of z, and $Z_{ij}$ is a function of r.

The interface zone ($0<z<\infty$) and the crack zone ($d<r<e$) is divided into 'n' and 'm' segments, respectively. These segment points along the interface and the crack surface are chosen as

$$\begin{aligned} \omega_i &= -\cos\left(\frac{i-1}{n_o\pi}\right)\frac{z_1}{2}, & i&=1,\ldots,n_o, \\ &= -\cos\left(\frac{i-1}{n_s\pi}\right)\frac{(z_2-z_1)}{2} + \frac{(z_2+z_1)}{2}, & i&=n_o+1,\ldots,n_o+n_s, \\ &= -\cos\left(\frac{i-1}{n_t\pi}\right)(z_3-z_2) + z_3, & i&=n_o+n_s+1,\ldots,n, \end{aligned} \tag{33.a}$$

$$\psi_i = x_i, \quad i=1,\ldots,m \tag{33.b}$$

where

$z_1$ = length of the open zone,

$z_2-z_1$ = length of the slip zone,

$z_3$ = maximum$[n^*+z_2,\ n^*(c-b)+z_2]$,

$n^*$ = a number chosen large enough to account for most of the stress changes along the interface,

$x_i$ = ith root of the mth order Legendre polynomial.

The above choice of segment points in equations (33) allows a concentration of segments near the transition points, such as, at the end of the open and slip zones, and at the transverse crack tip. The segments $n_o$, $n_s$, and $n_t$ are the number of segments in the open, slip and stick zones of the interface, respectively. Hence, the total number of segments at the interface(r=a) is

$$n = n_0 + n_s + n_t. \tag{34}$$

The collocation points along the interface and the crack length are chosen at the middle of the segment points as

$$\Omega_i = (\omega_i + \omega_{i+1})/2, \quad i=1,\ldots,n, \tag{35.a}$$

$$\zeta_i = (\psi_i + \psi_{i+1})/2, \quad i=1,\ldots,m. \tag{35.b}$$

There are (8n-4+2m) unknowns and one needs to set up the same number of equations. These are generated as follows.

1. The interface shear stresses $T_0$ and $T_1$, the interface normal stresses $S_0$ and $S_1$, and the crack slope function, G(r) are continuous at all points. These points include the segment points which give (2n+m-3) equations.

   Continuity of shear tractions along the interface at segment points $\omega_i$ gives

$$A_i + B_i\omega_{i+1} = A_{i+1} + B_{i+1}\omega_{i+1}, \quad i=1,\ldots,n-2, \tag{36.a}$$

$$A_{n-1} + B_{n-1}\omega_n = B_n/\omega_n^3, \tag{36.b}$$

   Continuity of normal tractions along the interface at segment points $\omega_i$ gives

$$C_i + D_i\omega_{i+1} = C_{i+1} + D_{i+1}\omega_{i+1}, \quad i=1,\ldots,n-2, \tag{36.c}$$

$$C_{n-1} + D_{n-1}\omega_n = D_n/\omega_n^3, \tag{36.d}$$

   Continuity of slope functions $\phi(r)$ along the transverse crack at segment points $\psi_i$ gives

$$U_i + V_i\psi_{i+1} = U_{i+1} + V_{i+1}\psi_{i+1}, \quad i=1,\ldots,m-1. \tag{36.e}$$

2. The continuity conditions (1) of shear and normal tractions at (r=a) give (4n-2) equations

$$A_i = P_i, \quad C_i = R_i, \quad i=1,\ldots,n-1, \tag{37.a}$$

$$B_i = Q_i, \quad D_i = S_i, \quad i=1,\ldots,n. \tag{37.b}$$

3. The open zone condition (2.a) of zero shear stress gives ($n_o$) equations as

$$A_i + B_i \Omega_i = 0, \; i=1,\dots,n_o. \tag{38}$$

4. The open zone condition (2.b) of zero normal tractions gives ($n_o$) equations as

$$C_i + D_i \Omega_i = -[\sigma_{rr}^{0T}(a,\Omega_i) + \sigma_{rr}^{0e}(a,\Omega_i)], \; i=1,\dots,n_o. \tag{39}$$

5. The Coulomb friction law in the slip zone condition (2.e) gives ($n_s$) equations as

$$A_i+B_i\Omega_i + \rho(C_i+D_i\Omega_i) = -\rho[\sigma_{rr}^{0T}(a,\Omega_I)+\sigma_{rr}^{0e}(a,\Omega_i)], \; i=n_o+1,\dots,n_o+n_s. \tag{40}$$

6. The radial displacement continuity conditions (2.d) and (2.h) in the slip and stick zone gives ($n_s$ + $n_t$) equations from equations (32.a) and (32.c), as

$$\begin{aligned}&\sum_{i=1}^{n-1} A_i X_{i1}(\Omega_j) + \sum_{i=1}^{n} B_i X_{i2}(\Omega_j) + \sum_{i=1}^{n-1} C_i X_{i3}(\Omega_j) + \sum_{i=1}^{n} D_i X_{i4}(\Omega_j)\\ &-\sum_{i=1}^{n-1} P_i Y_{i1}(\Omega_j) - \sum_{i=1}^{n} Q_i Y_{i2}(\Omega_j) - \sum_{i=1}^{n-1} R_i Y_{i3}(\Omega_j) - \sum_{i=1}^{n} S_i Y_{i4}(\Omega_j)\\ &-\sum_{i=1}^{m} U_i Y_{i5}(\Omega_j) - \sum_{i=1}^{m} V_i Y_{i6}(\Omega_j) = 0, \; j=n_o+1,\dots,n.\end{aligned} \tag{41}$$

7. The axial displacement continuity condition (2.i) in the stick zone gives ($n_t$) equations from equations (32.b) and (32.d) as

$$\begin{aligned}&\sum_{i=1}^{n-1} A_i X_{i5}(\Omega_j) + \sum_{i=1}^{n} B_i X_{i6}(\Omega_j) + \sum_{i=1}^{n-1} C_i X_{i7}(\Omega_j) + \sum_{i=1}^{n} D_i X_{i8}(\Omega_j)\\ &-\sum_{i=1}^{n-1} P_i Y_{i7}(\Omega_j) - \sum_{i=1}^{n} Q_i Y_{i8}(\Omega_j) - \sum_{i=1}^{n-1} R_i Y_{i9}(\Omega_j) - \sum_{i=1}^{n} S_i Y_{i10}(\Omega_j)\\ &-\sum_{i=1}^{m} U_i Y_{i11}(\Omega_j) - \sum_{i=1}^{m} V_i Y_{i12}(\Omega_j) = 0, \; j=n_o+n_s+1,\dots,n.\end{aligned} \tag{42}$$

8. The traction free crack surface condition (7) gives (m) equations from equation (32.e) as

$$\sum_{i=1}^{n-1} P_i Z_{i1}(\zeta_j) + \sum_{i=1}^{n} Q_i Z_{i2}(\zeta_j) + \sum_{i=1}^{n-1} R_i Z_{i3}(\zeta_j) + \sum_{i=1}^{n} S_i Z_{i4}(\zeta_j)$$
$$+ \sum_{i=1}^{m} U_i Z_{i5}(\zeta_j) + \sum_{i=1}^{m-1} V_i Z_{i6}(\zeta_j) = -[\sigma_{zz}^{1T}(\zeta_j,0) + \sigma_{zz}^{1\epsilon}(\zeta_j,0)], j=1,\ldots,m. \tag{43}$$

9. Since $u_z(c,z)$ is a constant as given by equation (3), the slope of the crack opening displacement $\frac{\partial}{\partial z}u_z(r,0)$ at the outer edge (r=e) of the hollow cylinder is zero. From equations (24) and (31.a) this gives

$$U_m + c\, V_m = 0. \tag{44}$$

The stress intensity factor (SIF) of the crack tip (r=d) is given by

$$K = \lim_{r \to d-} \sqrt{2(d-r)}\, \sigma_{zz}^{1}(r,0). \tag{45.a}$$

According to Gupta (1973), the SIF can be written as

$$K = \frac{\mu_1}{2(1-\nu_1)} \lim_{r \to d+} \sqrt{2(r-d)}\, \frac{\partial}{\partial r} u_z^1(r,0) = \frac{1}{\sqrt{2}}(U_1 + V_1 d). \tag{45.b}$$

The total number of equations (36)-(44) is (8n - 4 + 2m). These are solved simultaneously to calculate the unknown functions. One can then substitute these values in equation (29) to find the displacements and stresses at any point in the composite cylinder.

<u>Edge crack touching the slipping or open interface(d=a, e=c)</u>

The following steps are different for this case than given by equation (30-44).

1. Since the crack goes through the axial plane z=0, the force equilibrium is enforced in the axial direction in the composite cylinder. This is given by

$$-\int_0^\infty \sigma_{rz}^1(a,z)\ 2\pi b\ dz + [\sigma_{zz}^{1T}(r,\infty) + \sigma_{zz}^{1e}(r,\infty)]\ \pi(c^2-b^2) = 0 \qquad (46.a)$$

Substituting equation (30.c) in equation (46.a), we get

$$\sum_{j=1}^{n-1} 2\pi b(\omega_{j+1}-\omega_j)A_j + \sum_{j=1}^{n-1} \pi b(\omega_{j+1}^2-\omega_j^2)B_j + \frac{1}{2\omega_n^2}B_n \qquad (46.b)$$

$$= (\sigma_{zz}^{1T}(r,\infty)+\sigma_{zz}^{1e}(r,\infty)\pi(c^2-b^2)$$

The nth equation of equation (42) is replaced by equation (46.b).

2. The weight function of equation (31.b) is $\omega(r) = 1$ since the shear stress is symmetric and zero at (r=b, z=0). The absence of any singularities for relevant values of elastic moduli and friction coefficient, $\rho$ for ceramic matrix composites is discussed by Schweitert and Steif (1991). If the interface is open at z=0, singularities again do not exist in the slope function $\phi(r)$ (Lu and Erdogan, 1984).

3. Since the hollow cylinder can now have a rigid body displacement in the z direction, the first (n-1) equations of equation (42) are replaced by the continuity of the axial displacement differences at the interface as

$$u_z^0(a,\Omega_j) - u_z^0(a,\Omega_{j+1}) = u_z^1(a,\Omega_j) - u_z^1(a,\Omega_{j+1}),\ j=1,\ldots,n-1$$

which gives

$$\sum_{i=1}^{n-1} A_i X_{i5}(\Omega_j) + \sum_{i=1}^{n} B_i X_{i6}(\Omega_j) + \sum_{i=1}^{n-1} C_i X_{i7}(\Omega_j) + \sum_{i=1}^{n} D_i X_{i8}(\Omega_j)$$

$$- \sum_{i=1}^{n-1} A_i X_{i5}(\Omega_{j+1}) - \sum_{i=1}^{n} B_i X_{i6}(\Omega_{j+1}) - \sum_{i=1}^{n-1} C_i X_{i7}(\Omega_{j+1}) - \sum_{i=1}^{n} D_i X_{i8}(\Omega_{j+1})$$

$$
\begin{aligned}
&- \sum_{i=1}^{n-1} P_i Y_{i7}(\Omega_j) - \sum_{i=1}^{n} Q_i Y_{i8}(\Omega_j) - \sum_{i=1}^{n-1} R_i Y_{i9}(\Omega_j) - \sum_{i=1}^{n} S_i Y_{i10}(\Omega_j) \\
&- \sum_{i=1}^{m} U_i Y_{i11}(\Omega_j) - \sum_{i=1}^{m} V_i Y_{i12}(\Omega_j) \qquad (47) \\
&+ \sum_{i=1}^{n-1} P_i Y_{i7}(\Omega_{j+1}) + \sum_{i=1}^{n} Q_i Y_{i8}(\Omega_{j+1}) + \sum_{i=1}^{n-1} R_i Y_{i9}(\Omega_{j+1}) + \sum_{i=1}^{n} S_i Y_{i10}(\Omega_{j+1}) \\
&+ \sum_{i=1}^{m} U_i Y_{i11}(\Omega_{j+1}) + \sum_{i=1}^{m} V_i Y_{i12}(\Omega_{j+1}) = 0, \; j=n_o+n_s+1, \ldots, n-1.
\end{aligned}
$$

4. The shear stress at $z=z_1$ is exactly zero and is enforced by replacing the n^th equation of equation (41) as $\tau_{rz}(b, z_1) = 0$,

$$
A_{n_o+1} + B_{n_o+1}\, \omega_{n_o+1} = -\rho\; [\sigma_{rr}^{0T}(a, \omega_{n_o+1}) + \sigma_{rr}^{1e}(a, \omega_{n_0+1})] \qquad (48)
$$

For the above case, there are (8n - 4 + 2m) equations and equal number of unknowns. This system of linear equations is solved to get the unknowns.

The input remote strain $\epsilon_0$ may not satisfy all the inequality and sign conditions (2.c), (2.f), (2.g), (2.k) and (7). Hence, the strain $\epsilon_0$ is changed iteratively till it satisfies all the inequality conditions. It is quite possible that a small range of strain $\epsilon_0$ may satisfy all the equations including the inequality conditions. The correct $\epsilon_0$ is selected by the unique value where the shear stress at the slip-stick transition $(z=z_2)$ is smooth, that is

$$
\frac{d}{dz}\tau_{rz}^{1}(a, z)\,|_{z_2^-} = \frac{d}{dz}\tau_{rz}^{1}(a, z)\,|_{z_2^+},
$$

or

$$
B_{n_0+n_s} = B_{n_0+n_s+1}. \qquad (49)
$$

This smoothness of the shear stress at the slip-stick transition is implied by the asymptotic analysis of Dundurs and Comninou (1979) at the slip-stick transition in a frictional interface between two dissimilar elastic half-planes.

Closed crack in the matrix (d > a, e < c)

The following steps are different for this case than given by equations (30) - (44).

Equation (44) is replaced by the closed crack condition

$$\int_{d}^{e} \phi(r)\, dr = 0, \tag{50.a}$$

which gives

$$\sum_{i=1}^{m} U_i \int_{r_i}^{r_{i+1}} w(r)\, dr + \sum_{i=1}^{m} V_i \int_{r_i}^{r_{i+1}} r\, w(r)\, dr = 0,\ i=1,\ldots,m, \tag{50.b}$$

where

$$w(r) = \frac{1}{\sqrt{(r-d)\ (e-r)}}. \tag{50.c}$$

For this case, there are (8n - 4 + 2m) unknowns and an equal number of equations. Hence, the system of simultaneous linear equations can be solved.

For each of the above three cases, the solution can be substituted in equation (29) to get the stress and displacement field in the fiber and the matrix.

The computer program for this study is computationally very intensive. It takes about 15 hours of CPU time on a IBM 3090 computer for calculation of the slip lengths, and an extra 15 hours of CPU time for calculation of the critical stresses/displacements in the composite geometry. Hence, limited results are being presented in this paper.

We have, however, taken advantage of the method described in this paper to reutilize the intermediate results to solve the problem for different remote strain, $\epsilon_0$, the temperature change, $\Delta T$, friction coefficient, $\rho$, the coefficients of thermal expansion, $\alpha_0$ and $\alpha_1$. This is done as follows. The most computationally intensive part of the computer program are calculation of the elements of equations (32a-e) at the collocation points. However, once the (8n - 4 + 2m) equations are setup, the remote strain, $\epsilon_0$, the temperature change, $\Delta T$, friction coefficient, $\rho$, the coefficients of thermal expansion, $\alpha_0$ and $\alpha_1$ can be changed in these equations. These parameters do not change the coefficient matrix elements corresponding to equations (32). Other elements of the coefficient matrix and the right hand sides, which do need to be changed, require only a few seconds of computational time.

## RESULTS AND DISCUSSION

The numerical scheme presented in this study was tested as follows. Extensive convergence studies were done to find the number of collocation points to be used. In addition to numerical tests such as recovering applied stresses on the two bodies, comparisons with some exact models were also made. These tests included comparing the stress intensity factor for a small edge crack under uniform pressure, p which is $p\sqrt{2(e-d)}$ (Sneddon, 1951). For edge cracks comparable to the thickness, (c-b) of the hollow cylinder inside the matrix, the results for the stresses and the stress intensity factors matched within 5% with an independent computer program written for the problem solved by Wijeyewickrema and Keer (1991) using numerical methods enumerated by Kaw and Pagano (1993).

The SiC/CAS material system with the following elastic properties are used to discuss the results.

Silicon Carbide Fiber:

$E_0$ = 207 *MPa*

$\nu_0$ = 0.25

Calcium Alumino Silicate Glass Matrix:

$E_1$ = 98 *MPa*

$\nu_1$ = 0.25

$V_f$ = 0.4.

Only cases with slip and no open zone are considered in this study. This is because for cases where both types of zones are possible, there are two variables, the length of the open zone and the length of the slip zone, which need to be found iteratively. Finding these two variables is computationally tractable with the current computer program but would require prohibitive amount of computational time.

Figure 3 shows the normalized slip length, $L_s/a$ as a function of the ratio of the matrix axial stress to interfacial radial pressure (ARS),

$$ARS=[\sigma^{1T}_{zz}(r,\infty)+\sigma^{1e}_{zz}(r,\infty)]\,/\,[\sigma^{1T}_{rr}(a,\infty)+\sigma^{1e}_{rr}(a,\infty)] \qquad (51)$$

for constant coefficient of friction. The reason for choosing the abscissa as the ratio of the stresses, ARS is because any combination of temperature change, remote axial strain and linear coefficients of thermal expansion which result in the same stress ratio (ARS) will result in the same slip length.

In Figure 3, the slip length increases linearly with the stress ratio, ARS and decrease linearly with the coefficient of friction. Note that the slip length was also found (not shown) to be a linear function of the remote axial strain since the stress $\sigma_{rr}^{1e}(a,z)=0$ for equal Poisson's ratio of the fiber and the matrix.

In Figure 3, the value of ARS=1.618 corresponds to zero remote axial strain. It is important to note that at this value of ARS, the interface damage is assumed to have already taken place only due to the residual tensile stresses in the matrix.

The slip lengths obtained from our model are compared with those obtained by using a equivalent Gu and Mangonon (1992) type model. Figure 4 shows the ratio of the slip lengths obtained from Gu and Mangonon's (1992) type model and the present model as a function of the stress ratio, ARS for constant coefficients of friction. The ratio of slip length approaches one as ARS increases, which may be an indication that for large slip lengths, the predictions from the two models are same.

The following properties are assumed additionally to discuss the following results.

$\alpha_0 = 3.5 \; x \, 10^{-05} \; m/m/^oC$

$\alpha_1 = 6.5 \; x \, 10^{-05} \; m/m/^oC$

$\Delta T = -1000^oC.$

The main assumptions in the Gu and Mangonon's (1992) model are that the axial stresses ( $\sigma_{zz}$) are independent of the radial co-ordinate. However, if one looks at Figure 5, where the fiber axial stresses at the crack plane (z=0) are plotted as a function of the normalized radial co-ordinate, r/a, one can see that this assumption is not valid. The stresses, however, remain fairly constant far away from the interface and far away from the crack plane. One may, however, point out that as the slip length increases, the stress concentration factor in the fiber, SCF defined by

$$SCF=M_4^0(a-,0)/\sigma_{zz}^{0e}(r,\infty) \qquad (52)$$

decreases and may be an indication that for large slip lengths, Gu and Mangonon's (1992) model and the present model give similar results.

In Figure 6, the interfacial radial and shear stresses are plotted as a function of the normalized axial location, z/a for constant slip lengths (or constant remote axial strain). The radial stress increases rapidly to the remote radial stress at the interface. Note the small effect of the increasing slip length (increasing remote strain) on the interfacial radial stresses.

The interfacial shear stress in the slip zone follow the same pattern as the radial stresses, since they are linearly related in the slip zone. At the end of the slip zone, the shear stress decays rapidly to zero. Note that the maximum interfacial shear stress does not change with increasing slip length (increasing remote strain).

The conclusions from Figure 6 show that the assumption of constant shear stress for a low friction coefficient used in other models (Wijeyewickrama and Keer, 1993) may be valid. However, one should note that a constant shear stress assumption gives logarithmically singular fiber axial stresses at the crack tip, (r=a-,z=0), while the Coulomb friction law gives large but finite fiber axial stresses at the crack tip.

## CONCLUSIONS

The main conclusions of this study made for a typical SiC/CAS system under a negative temperature change and a remote axial strain are

1. The length of the slip zone increases linearly with increasing remote axial strain and decreases linearly with low coefficients of friction.

2. The stress concentration factor in the fiber at the crack tip decreases with an increase in the remote axial strain.

3. The interfacial radial and shear stresses for low coefficients of friction are nearly independent of the remote axial strain. Moreover, these stresses are fairly constant in the slip zone.

## REFERENCES

Abramowitz, M. and Stegun, I.A. (1970), *Handbook of Mathematical Functions*, Dover, New York.

Aksel, B., Hui, C. and Lagoudas, D.C. (1991). Effects of a frictional interface on the load diffusion from a broken filament embedded in an elastic medium. *Int. J. Solids Structures.* **27**, 833-847.

Dundurs, J. and Comninou, M. (1979). Some consequences of the inequality conditions in contact and crack problems. *J. Elasticity* 9, 71-82.

Erdol, R. and Erdogan, F. (1978). A thick-walled cylinder with an axisymmetric internal of edge crack. *ASME J. Appl. Mech.* 45, 281-286.

Gu, L. and Mangonon, P. L. (1992). Mechanical characteristics of fiber-reinforced brittle matrix composite, case II: non-zero radial stress at matrix external surface. Constitutive behavior of high-temperature Composites, MD-Vol 40. ASME, pp. 151-161.

Gupta, G.D. (1973). A layered composite with a broken laminate. *Int. J. Solids Structures.* **9**, 1141-1154.

Kaw, A.K. and Pagano, N.J. (1993). Axisymmetric thermoelastic response of a composite cylinder containing an annular matrix crack. *J. Comp. Mat.* **27**, 540-571.

Lu, M. and Erdogan, F. (1984). Stress intensity factors in two bonded elastic layers containing cracks perpendicular to and on the interface. *Eng. Frac. Mech.* **18**, 491-506.

Steif, P. S. (1984). Stiffness reduction due to fiber breakage. *J. Comp. Mat.* **18**, 153-172.

Schweitert, H.R. and Steif, P.S. (1991). Analysis of a broken fiber in a weakly bonded composite. *Int. J. Solids Structures.* **28**, 283-297.

Wijeyewickrema, A.C. and Keer, L.M. (1991). Matrix cracking in a fiber reinforced composite with slip at the fiber-matrix interface. *Int. J. Solids Structures.* **30**, 91-113.

Wijeyewickrema, A.C. and Keer, L.M. (1993). Matrix fracture in brittle matrix fiber-reinforced composites. *Int. J. Solids Structures.* **28**, 43-65.

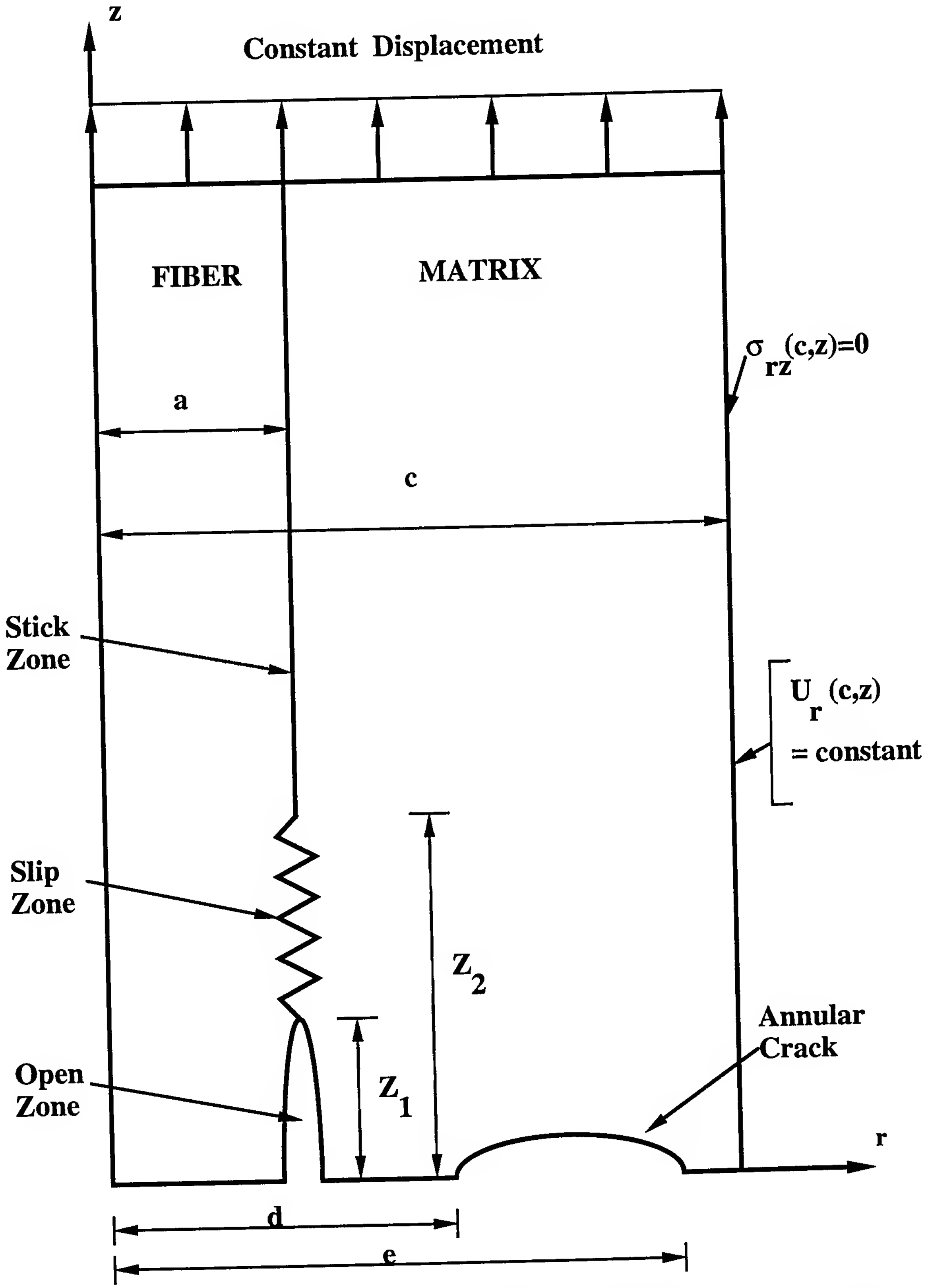


Figure 1. Schematic of a Representative Volume Element of a Brittle Matrix Composite with Frictional Interfaces and Matrix Cracking

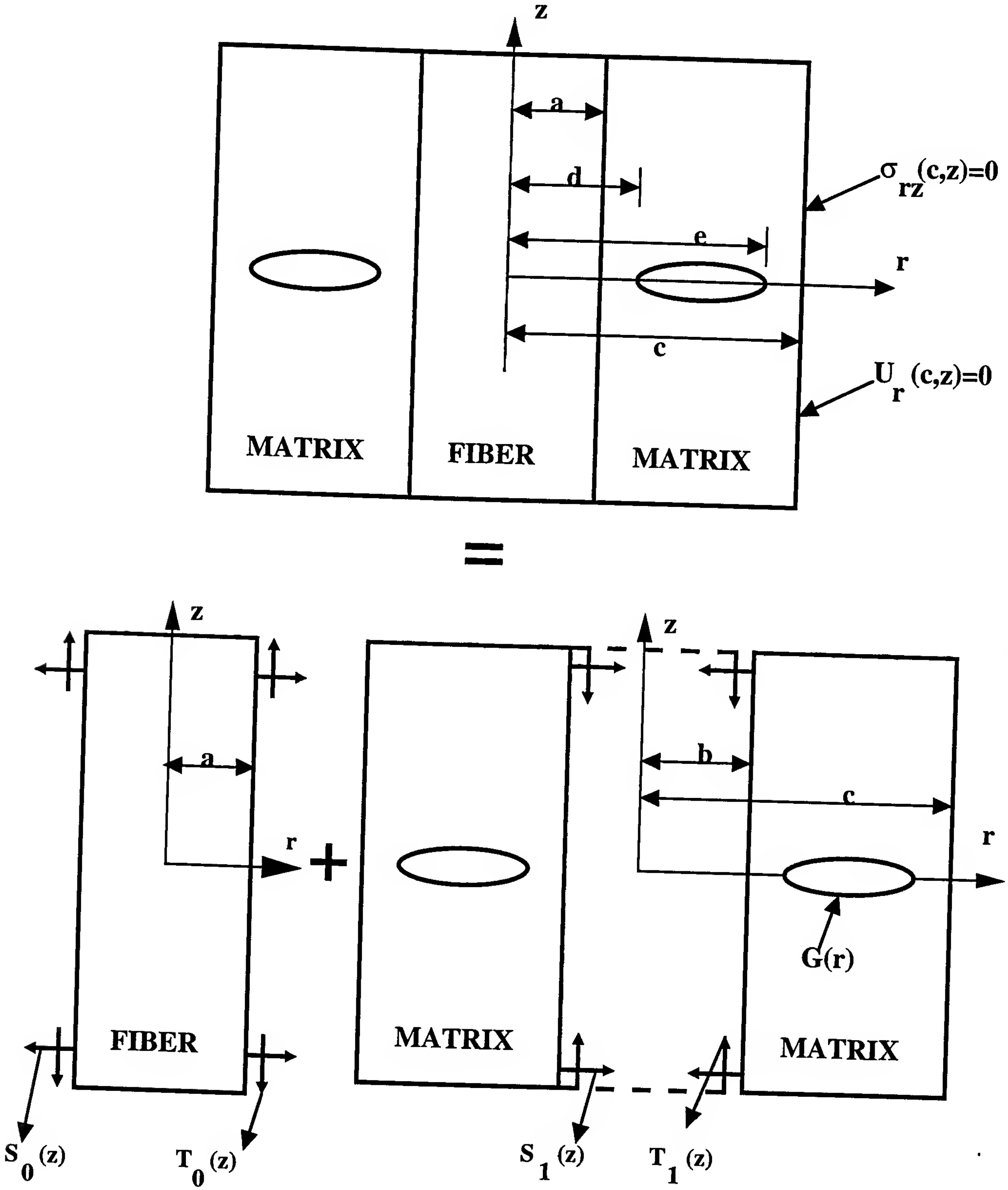


Figure 2. Schematic of Free-Body Diagrams of Fiber and Matrix with Unknown Functions

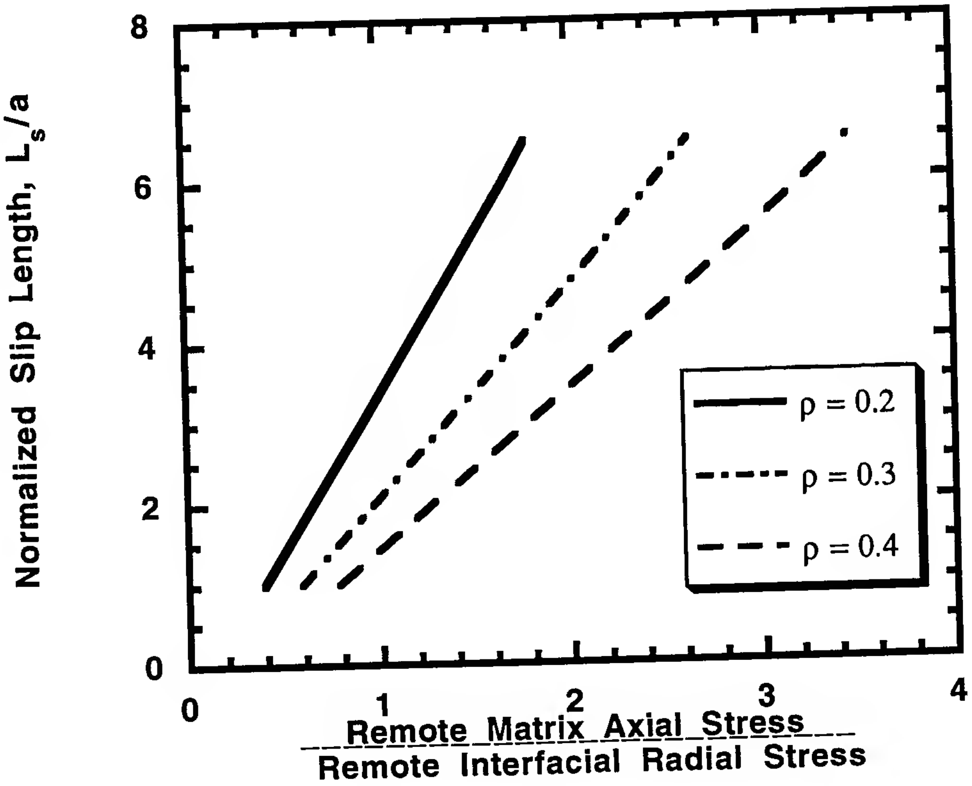


Figure 3. Normalized Slip Length as a Function of the Ratio of Remote Matrix Axial Stress to Remote Interfacial Radial Stress for Constant Coefficients of Friction.

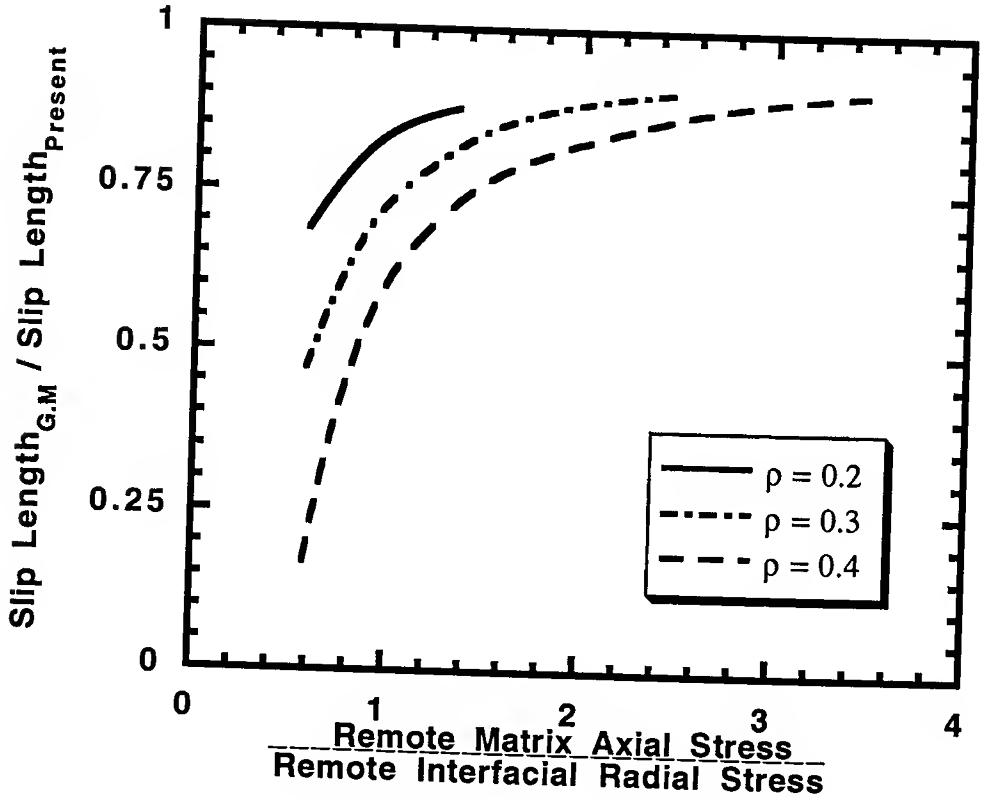


Figure 4. Ratio of Slip Length of Gu and Mangonon Type Model to Present Model as a Function of Ratio of Remote Matrix Axial Stress to Remote Interfacial Radial Stress for Constant Coefficients of Friction.

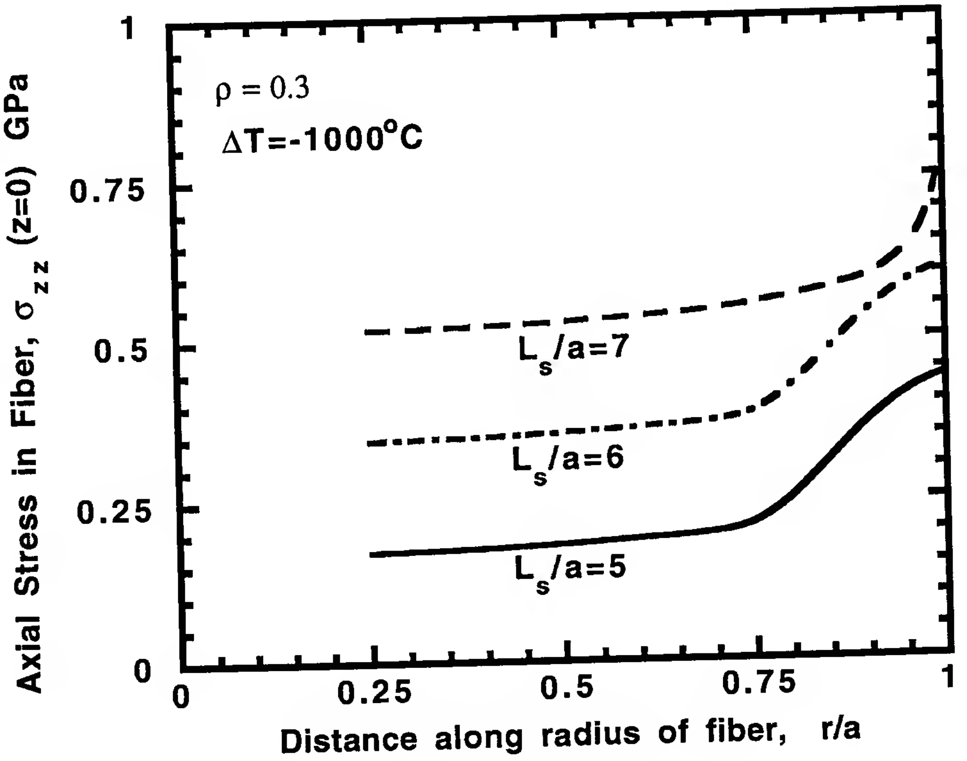


Figure 5. Fiber Axial Stresses at Crack Plane as a Function of Radial Location for Constant Slip Lengths

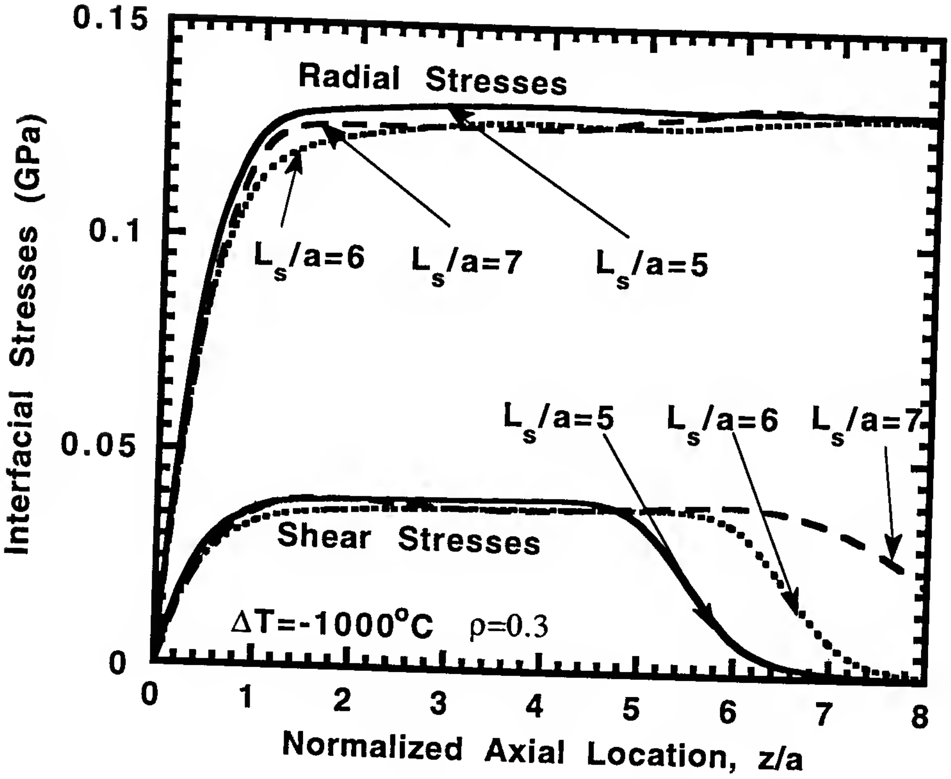


Figure 6. Interfacial Stresses as a Function of Normalized Axial Location for Constant Slip Lengths.

# COMBINED EFFECTS OF GRADED AND SETBACK LAYERS ON THE AlGaAs/GaAs HBT CURRENT-VOLTAGE CHARACTERISTICS

J. J. Liou
Associate Professor
Electrical & Computer Engineering Dept.
University of Central Florida, Orlando, FL 32816

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

University of Central Florida

December 1993

# COMBINED EFFECTS OF GRADED AND SETBACK LAYERS ON THE AlGaAs/GaAs HBT CURRENT-VOLTAGE CHARACTERISTICS

J. J. Liou
Associate Professor
Electrical & Computer Engineering Dept.
University of Central Florida, Orlando, FL 32816

## Abstract

The combined effects of graded and setback layers ($W_G$ and $W_I$) on the AlGaAs/GaAs heterojunction bipolar transistor (HBT) d.c. performance are investigated, and an analytical model which can describe the behavior of such HBTs is presented. The HBT base and collector currents accounting for the variation of the conduction and valence bands due to the presence of $W_G$ and $W_I$ are also calculated. It is shown that including $W_G$ and $W_I$ actually degrades the HBT current gain at low current levels. The current gain at high current levels, on the other hand, can be enhanced if $W_I$ = 150 Å and $0 \leq W_G \leq 300$ Å or $W_I = 0$ and 150 Å $\leq W_G \leq 300$ Å are used. The model predictions compare favorably with results calculated from a numerical model.

# COMBINED EFFECTS OF GRADED AND SETBACK LAYERS ON THE AlGaAs/GaAs HBT CURRENT-VOLTAGE CHARACTERISTICS

J. J. Liou

## 1. INTRODUCTION

The superior performance of the AlGaAs/GaAs heterojunction bipolar transistor (HBT) results directly from the valence band discontinuity $\Delta E_V$ at the hetero-interface, which arises from the wider bandgap in the emitter than base [1]. The benefit of having $\Delta E_V$ is two-fold. First, it allows a high base doping concentration, which reduces the base series resistance and thus the a.c., d.c., and transient emitter crowding, while maintaining an acceptable current gain. Second, because of the high base doping concentration, the base region can be made very thin, thus reducing the base transit time and increasing the cutoff frequency [2].

The conduction band discontinuity $\Delta E_C$ (spike) at the hetero-interface, on the other hand, is not as desirable as $\Delta E_V$. This is because the spike necessitates the free carriers in the heterojunction to transport by means of thermionic and tunneling mechanisms [3-4]. This impedes the free-carrier injection from the emitter to base and decreases the collector current. The problem is often overcome by inserting a thin layer (graded layer) before the hetero-interface in the which the Al mole fraction is graded linearly and/or a thin undoped GaAs layer (setback layer or spacer) after the hetero-interface. The junction grading will lower or even remove the spike, and thus reducing the importance of thermionic and tunneling and making the injection more efficient [5]. On the other hand, inserting a setback layer does not alter the spike, but rather decreases the barrier potential before the spike. This also makes the thermionic and tunneling less prominent and improves the injection efficiency [6]. Another advantage of the setback layer is that it can prevent impurity out-diffusion from the heavily doped base to emitter.

Most HBT models reported in the literature focus on the graded HBTs. Boundary conditions for the excess carrier concentrations at the space-charge region edges of the graded HBT were derived by Lundstrom [7]. These boundary conditions have been used to develop the thermionic-field-diffusion model including both thermionic and tunneling mechanisms [8]. A detailed analytic study on the space-charge region recombination current derived from the charge-control approach was presented [9]. Numerical solutions to the current transport in the graded junction are also available [10-11].

Studies on the effect of setback layer on the HBT current transport have been limited in the past. Such an effect was studied analytically in [6] by solving the Poisson equation accounting for the boundary conditions associated with the setback layer. A numerical model reported in [12] is also applicable, in which the Poisson and continuity equations are calculated accounting for the nonuniform spatial band distribution as well as carrier degeneracy. To the best of our knowledge, an analytical HBT model which treats comprehensively the graded and setback layers is yet absent.

This paper develops an analytical HBT model including the combined effects of graded and setback layers. The variation of the conduction and valence bands, as well as the base and collector currents, as functions of the graded and setback layer thicknesses will be studied in detail. Results will be presented for several graded- and setback-layer thickness combinations, and compared with those calculated from a numerical model.

## 2. MODEL DEVELOPMENT

### *2.1 Barrier Potentials and Space-Charge-Region Thicknesses*

Consider an N/p$^+$/n AlGaAs/GaAs HBT shown in Fig. 1. The position-dependent dielectric permittivity $\varepsilon_g(x)$ in the linearly graded layer ($-W_G \le x \le 0$) is given by

$$\varepsilon_g(x) = -(\varepsilon_E - \varepsilon_B)x/W_G + \varepsilon_B \qquad (1)$$

where $\varepsilon_E$ and $\varepsilon_B$ are the dielectric permittivities in the emitter and base, respectively, and $W_G$ is the graded layer thickness. The one-dimensional Poisson equation in the graded layer is

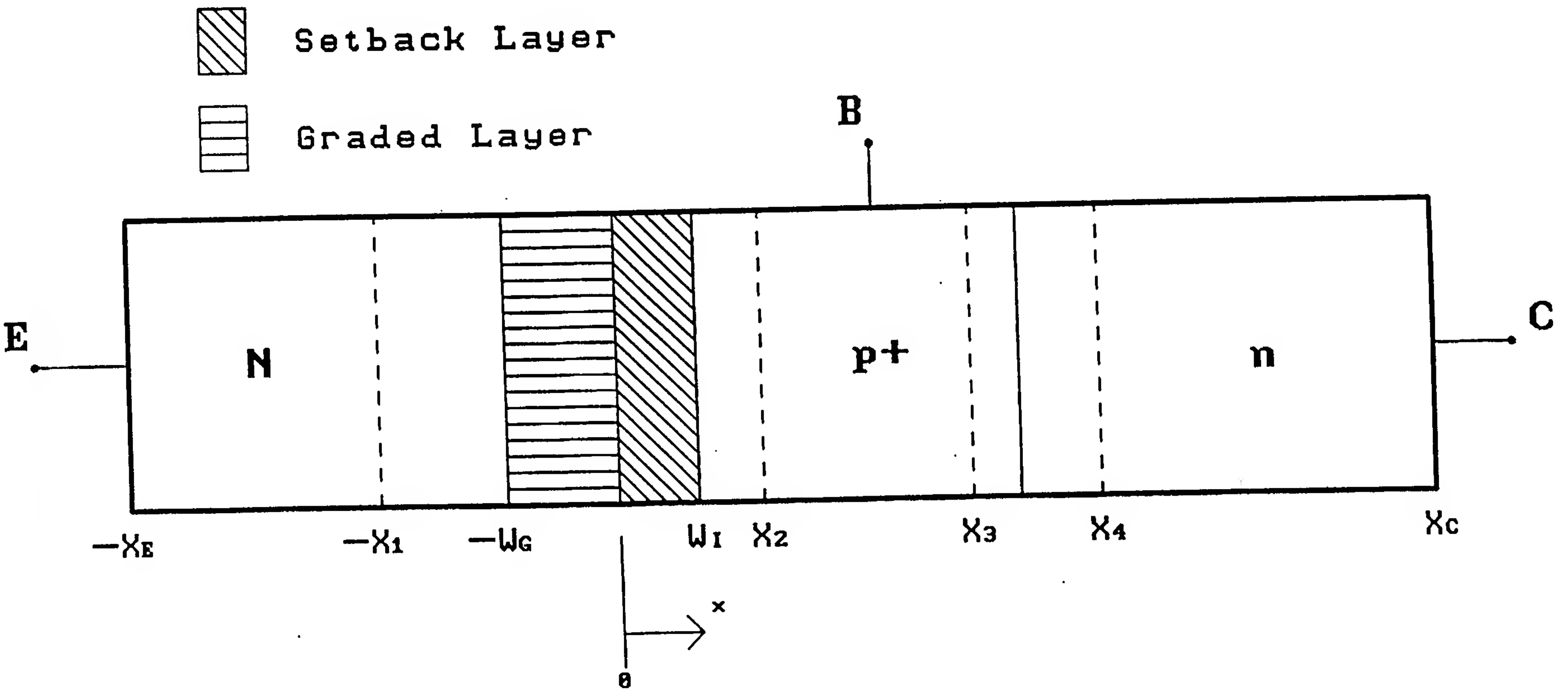


Fig. 1 The HBT structure including graded and setback layers.

$$d^2V/dx^2 = -\{\rho/\varepsilon_g(x) - (dV/dx)(\varepsilon_E - \varepsilon_B)/[W_G\varepsilon_g(x)]\} \tag{2}$$

where V is the electrostatic potential, and $\rho$ is the charge density. Employing the conventional depletion approximation and integrating the equation once, we obtain the electric field $\xi_g(x)$ in the graded layer as

$$\xi_g(x) = -dV/dx = (qN_Ex - C')/\varepsilon_g(x) \tag{3}$$

Here $N_E$ is the emitter doping concentration and C′ is a constant which needs to be determined from the boundary condition. Similarly, we can derive the electric fields $\xi_1(x)$ in the space charge region (SCR) on the emitter side, $\xi_s(x)$ in the setback layer, and $\xi_2(x)$ in the SCR on the base side:

$$\xi_1(x) = (qN_E/\varepsilon_E)(x + X_1) \quad \text{for } -X_1 < x < -W_G \tag{4}$$

$$\xi_s(x) = (qN_B/\varepsilon_B)(X_2 - W_I) \quad \text{for } 0 < x < W_I \tag{5}$$

$$\xi_2(x) = (qN_B/\varepsilon_B)(X_2 - x) \quad \text{for } W_I < x < X_2 \tag{6}$$

where $N_B$ is the base doping concentration, $W_I$ is the thickness of the setback layer, and $X_1$ and $X_2$ are the thicknesses of the SCR on emitter and base sides, respectively (see Fig. 1). Since the flux density at $x = -W_G$ is continuous, constant C′ in (3) can be obtained by using the boundary condition $\xi_g(-W_G) = \xi_1(-W_G)$:

$$\xi_g(x) = [qN_E/\varepsilon_g(x)](x + X_1) \tag{7}$$

By choosing $x = -X_1$ as the reference (zero point), the corresponding electrostatic potential V(x) can be calculated by integrating $\xi(x)$ over its boundary. Thus

$$V_1(x) = 0.5(qN_E/\varepsilon_E)(x + X_1)^2 \tag{8}$$

$$V_g(x) = V_1(-W_G) + qN_EW_G(x + W_G)/(\varepsilon_B - \varepsilon_E) + qN_EW_G(X_1 - W_G\varepsilon_E)/(\varepsilon_B - \varepsilon_E)^2\ln\{[(\varepsilon_B - \varepsilon_E)x + W_G\varepsilon_B]/(W_G\varepsilon_E)\} \tag{9}$$

$$V_s(x) = V_g(0) + (qN_B/\varepsilon_B)(X_2 - W_I)x \tag{10}$$

$$V_2(x) = V_s(W_I) + (qN_B/\varepsilon_B)(X_2x - x^2/2 - X_2W_I + W_I^2/2) \tag{11}$$

where

$$V_1(-W_G) = 0.5(qN_E/\varepsilon_E)(X_1 - W_G)^2 \tag{12}$$

$$V_g(0) = 0.5(qN_E/\varepsilon_E)(x + X_1)^2 + [qN_EW_G/(\varepsilon_B - \varepsilon_E)]\{x + W_G + [X_1 - W_G\varepsilon_E/(\varepsilon_B - \varepsilon_E)]\ln(\varepsilon_B/\varepsilon_E)\} \tag{13}$$

$$V_s(W_I) = V_g(0) + (qN_B/\varepsilon_B)(X_2 - W_I)W_I \tag{14}$$

Also, the total electrostatic potential $V_2(X_2)$ across the SCR should be equal to $V_{bi,BE} - V_{BE}$, where $V_{BE}$ is the applied base-emitter voltage and $V_{bi,BE}$ is the junction built-in potential [13]. Thus

$$V_{bi,BE} - V_{BE} = V_s(W_I) + 0.5(qN_B/\varepsilon_B)(X_2 - W_I)^2 \tag{15}$$

The SCR thickness $X_1$ can be found from (15) and $X_2 = (N_B/N_E)X_1 + W_I$ (derived from charge neutrality in the entire SCR):

$$X_1 = -0.5B/A - 0.5(B^2 - 4AC)^{0.5}/A \tag{16}$$

where:

$$A = 0.5qN_E/\varepsilon_E + 0.5qN_E^2/(\varepsilon_E N_B) \tag{17}$$

$$B = -qN_E W_G/\varepsilon_E - [qN_E W_G/(\varepsilon_B - \varepsilon_E)]\ln(\varepsilon_B/\varepsilon_E) + qN_E W_I/\varepsilon_E \tag{18}$$

$$C = -qN_E W_G^2/(\varepsilon_B - \varepsilon_E) + [qN_E W_G^2 \varepsilon_B/(\varepsilon_B - \varepsilon_E)^2]\ln(\varepsilon_B/\varepsilon_E) + 0.5qN_E W_G^2/\varepsilon_E \tag{19}$$

The barrier potentials for the conduction band edge $E_C$ in the SCR are defined as

$$V_{B1} = V_1(-W_G); \quad V_{BGC} = -\Delta E_C/q + V_g(0) - V_1(-W_G);$$

$$V_{BS} = V_s(W_I) - V_g(0); \quad V_{B2} = V_2(X_2) - V_s(W_I) \tag{20}$$

where $\Delta E_C$ is the conduction band discontinuity. Since the valence band edge $E_V$ is in parallel with $E_C$ except for that in the graded layer, the barrier potentials for the valance band are the same as that given in (20) provided $V_{BGC}$ is changed to

$$V_{BGV} = \Delta E_V/q + V_g(0) - V_1(-W_G) \tag{21}$$

$\Delta E_V$ is the valence band discontinuity. Note that all barrier potentials have positive values except for $V_{BGC}$, the value of which can be positive or negative depending on the applied voltage and thickness of the graded layer. A positive value indicates that $E_C$ have a positive slope in the graded layer, and vice versa if the value is negative.

### *2.2 Collector and Base Currents*

Following the thermionic-field-diffusion approach [7-8], the electron current density $J_n$ across the spike (located at $x = -W_G$) is the difference between two opposing fluxes:

$$J_n(-W_G) = qv_n\gamma[n(-W_G^-) - n(-W_G^+)] \qquad (22)$$

where $v_n$ is the electron thermal velocity, $\gamma$ is the tunneling coefficient [8], and n is the electron concentration. It should be mentioned that $\gamma$ depends strongly on $V_{B1}$ and $V_{BGC}$ if $V_{BGC}$ is negative and that $\gamma = 1$ (no tunneling) and the conventional drift-diffusion model applies if $V_{BGC}$ is positive. In (22),

$$n(-W_G^-) = N_E\exp(-V_{B1}/V_T) \quad \text{and} \quad n(-W_G^+) = n(X_2)\exp[(V_{B2} + V_{BS} + V_{BGC})/V_T] \qquad (23)$$

At this point, the only unknown parameter is $n(X_2)$, which can be found using the relation:

$$J_n(-W_G) = J_{SCRG} + J_{SCRS} + J_{SCR2} + J_n(X_2) \qquad (24)$$

where $J_{SCRG}$, $J_{SCRS}$, and $J_{SCR2}$ are the recombination current densities in the graded layer, setback layer, and space-charge layer in the base region, respectively (the models for $J_{SCRG}$, $J_{SCRS}$, and $J_{SCR2}$ will be developed in the next section), and $J_n(X_2)$ is the diffusion-only current in the quasi-neutral base (QNB). For a very thin base,

$$J_C \approx J_n(X_2) = qD_n n(X_2)/(W_B + D_n/v_{sat}) \qquad (25)$$

where $J_C$ is the collector current density, $D_n$ is the electron diffusion coefficient in the QNB, $W_B = X_3 - X_2$ (Fig. 1) is the QNB thickness, and $v_{sat}$ (= $10^7$ cm/sec) is the saturation drift velocity caused by the high field in the base-collector junction. An empirical expression can be used to describe $D_n$ [14]:

$$D_n = V_T[7200/(1 + 5.5x10^{-17}N_B)^{0.233}] \qquad (26)$$

Combining (22)-(25) and solving for $n(X_2)$, we obtain

$$n(X_2) = [qv_n\gamma N_E\exp(-V_{B1}/V_T) - J_{SCR2} - J_{SCRS} - J_{SCRG}]/\eta \tag{27}$$

where $\eta = qD_n/(W_B + D_n/v_{sat}) + qv_n\gamma\exp[(V_{B2} + V_{BS} + V_{BGC})/V_T]$.

The components of the base current density $J_B$ of the HBT include 1) injection of hole current density $J_{RE}$ from the base to emitter; 2) electron-hole recombination current density $J_{RB}$ in the quasi-neutral base; 3) electron-hole recombination current density $J_{SCR}$ in the emitter-base space-charge layer; and 4) electron-hole recombination current density $J_{RS}$ at the emitter and base surfaces.

The hole injection current can be modeled using the conventional diffusion-current only approximation:

$$J_{RE} = qD_pN_B\exp[-(V_{B1} + V_{B2} + V_{BS} + V_{BGV})/V_T]/W_E \tag{28}$$

where $D_p$ is the hole diffusion coefficient in the emitter and $W_E = X_E - X_1$ is the thickness of the quasi-neutral emitter. The doping-concentration dependent $D_p$ is given by [14]

$$D_p = V_T[380/(1 + 3.2x10^{-17}N_E)^{0.2666}] \tag{29}$$

The recombination current density in the quasi-neutral base is

$$J_{RB} = J_n(X_2)(1 - \alpha_B) \tag{30}$$

where $\alpha_B$ is the base transport factor:

$$\alpha_B = 1/\cosh[W_B/(D_n\tau_n)^{0.5}] \tag{31}$$

$\tau_n$ is the electron lifetime in the base ($\tau_n$ = 1 nsec is used in calculations).

Note that $\alpha_B$ approaches unity for a very thin base.

$J_{SCR}$ consists of four recombination current densities occurred in the emitter-side of the space-charge layer ($J_{SCR1}$), in the graded layer ($J_{SCRG}$), in the setback layer ($J_{SCRS}$), and in the base-side of the space-charge layer ($J_{SCR2}$). Based on the assumption that the Shockley-Read-Hall statistics is the dominant recombination process, then these current densities are proportional to the intrinsic free-carrier concentration and $\exp(V_{BE}/2V_T)$:

$$J_{SCR} = J_{SCR1} + J_{SCRG} + J_{SCRS} + J_{SCR2} = (n_{iE} + n_{iG} + n_{is} + n_{iB})B^{*}\exp(V_{BE}/2V_T) \qquad (32)$$

where $n_i$ is the intrinsic carrier concentration and $B^*$ is an empirical parameter that relates to the trapping density in the space-charge region.

The surface current density is influenced strongly by the fabrication process. It includes electron-hole recombination taking place at the emitter as well as base surfaces and can be empirically modeled as [15]

$$J_{RS} = J^{*}\exp(V_{BE}/nV_T) \qquad (33)$$

Here the ideality factor n is close to unity, and $J^*$ is the empirical parameter that characterized the recombination current at the emitter and base surfaces, which is a function of the surface recombination velocity S, the value of which depends strongly on the surface states and the location of the Fermi level pinned at the surface ($S = 10^6$ cm/sec is used in our calculations).

## 3. RESULTS AND DISCUSSIONS

For illustrations, we consider a typical HBT makeup having $5x10^{17}$, $10^{19}$, and $10^{16}$ $cm^{-3}$ emitter, base, and collector doping concentrations, and 1700, 1000, and 5000 Å emitter, base, and collector layer thicknesses, respectively. Also, the HBT has graded and setback layer thicknesses ranging from 0 to 300 Å. Since the results for HBTs having only the graded layer and only the setback layer have been reported previously [3-11], our emphasis here will be placed on the combined effects of the two layers on the HBT performance.

Fig. 2 illustrates the spatial variation of conduction and valence band edges of an HBT having $W_I = W_G = 150$ Å for three different $V_{BE}$. It is shown that while $E_V$ in the junction increases monotonically versus the position for all voltages, the slope of $E_C$ in the graded layer actually becomes negative as $V_{BE}$ is increased. Associated with such a negative slope is an "alloy" junction barrier which is formed at the vicinity of $x = -W_G$. The same results were also obtained in numerical simulation by Frank and Tiwari [16]. As will be shown later, the formation of alloy junction barrier increases the importance of the thermionic and tunneling mechanisms and subsequently reduces the collector current at large $V_{BE}$. Fig. 3 compares the conduction band edges of the same HBT calculated from the present model and obtained from a numerical model [12]. Good agreement is found between the two models.

We next study the collector and base currents, and their voltage dependencies are plotted in Figs. 4 and 5 respectively. The results suggest that, except for very large $V_{BE}$ at which the alloy junction barrier exists, both $J_C$ and $J_B$ increase considerably when $W_I$ and $W_G$ increase from 0 to 150 Å but change only slightly when $W_I$ and $W_G$ increase from 150 to 300 Å. The $J_C$ increase results from the removal of $\Delta E_V$ which reduces the importance of thermionic and tunneling mechanisms. On the other hand, the $J_B$ increase is due to the fact that the presence of $W_I$ and $W_G$ widens the space-charge layer and thus enhances electron-hole recombination in the region. Again the results calculated from the present model compare favorably with those obtained from the numerical model [12]. Note that when $W_I$ and $W_G$ are present, $J_C$ saturates at a smaller $V_{BE}$ (Fig. 4). This stems from the fact that the free-carrier transport in such HBTs is hinder by the alloy junction barrier formed at large $V_{BE}$ (see Fig. 3).

To better compare the currents in HBTs with $W_I$ and $W_G$ and without $W_I$ and $W_G$ (abrupt HBT), the currents in HBTs with $W_I$ and $W_G$ are normalized by those in abrupt HBTs. These normalized $J_C$ and $J_B$ are shown in Figs. 6 and 7, respectively. The results in Fig. 6 indicate that the HBT with $W_I = W_G = 150$ Å has the highest $J_C$ at relatively low $V_{BE}$, but all four HBTs have similar $J_C$ at

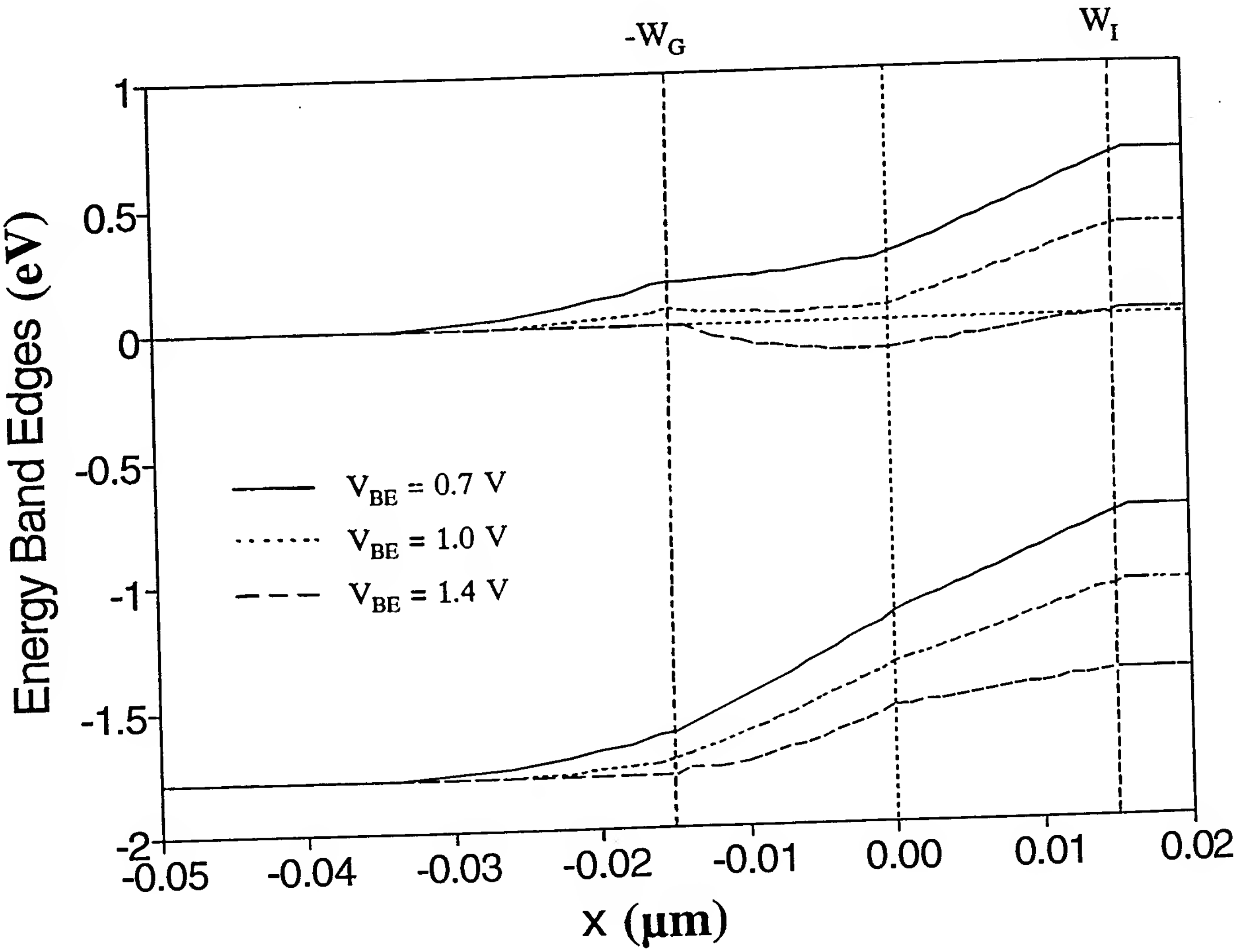


Fig. 2 Conduction and valence band edges calculated from the present model for an HBT with $W_I = W_G = 150$ Å.

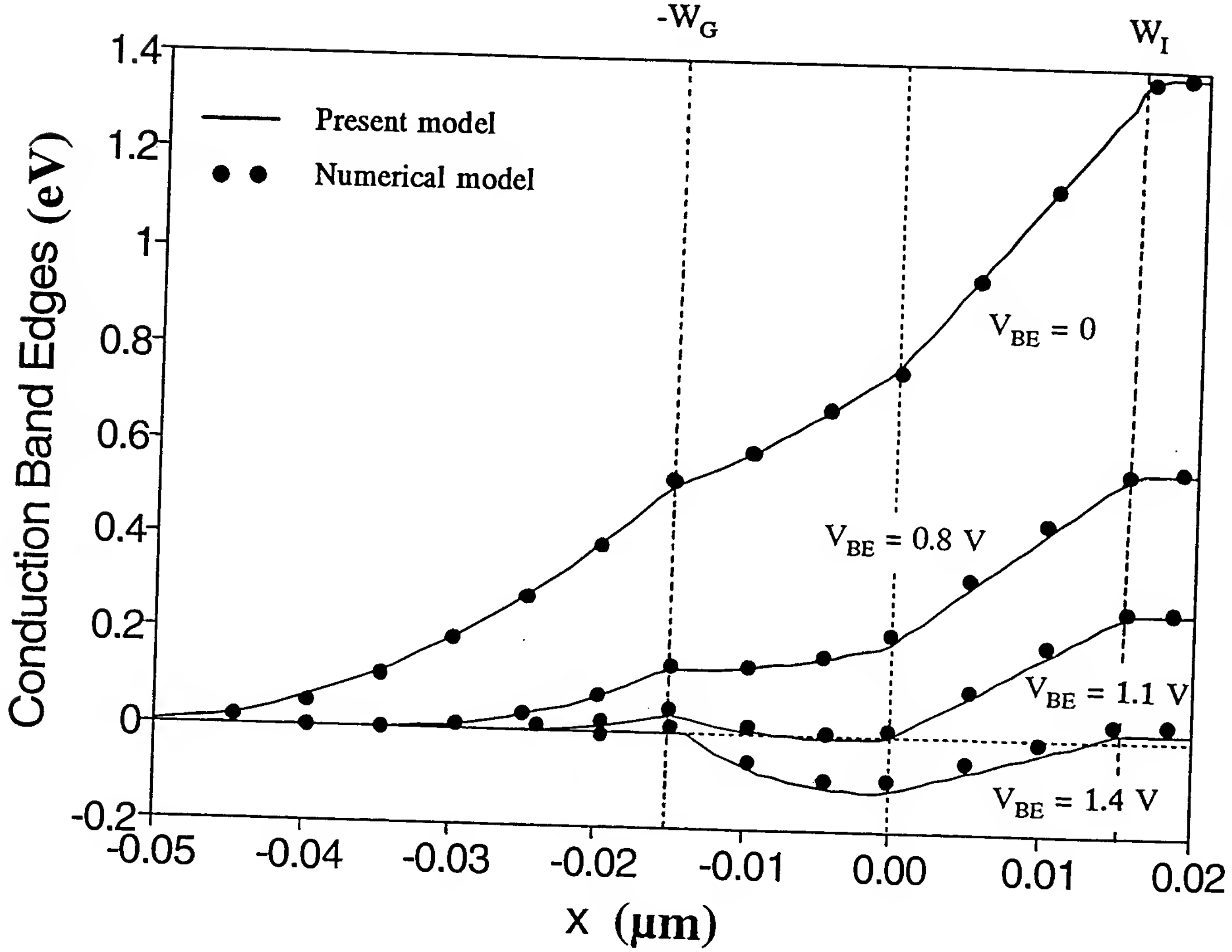


Fig. 3 Comparison of the conduction band edge calculated from the present model and numerical model at four different voltages.

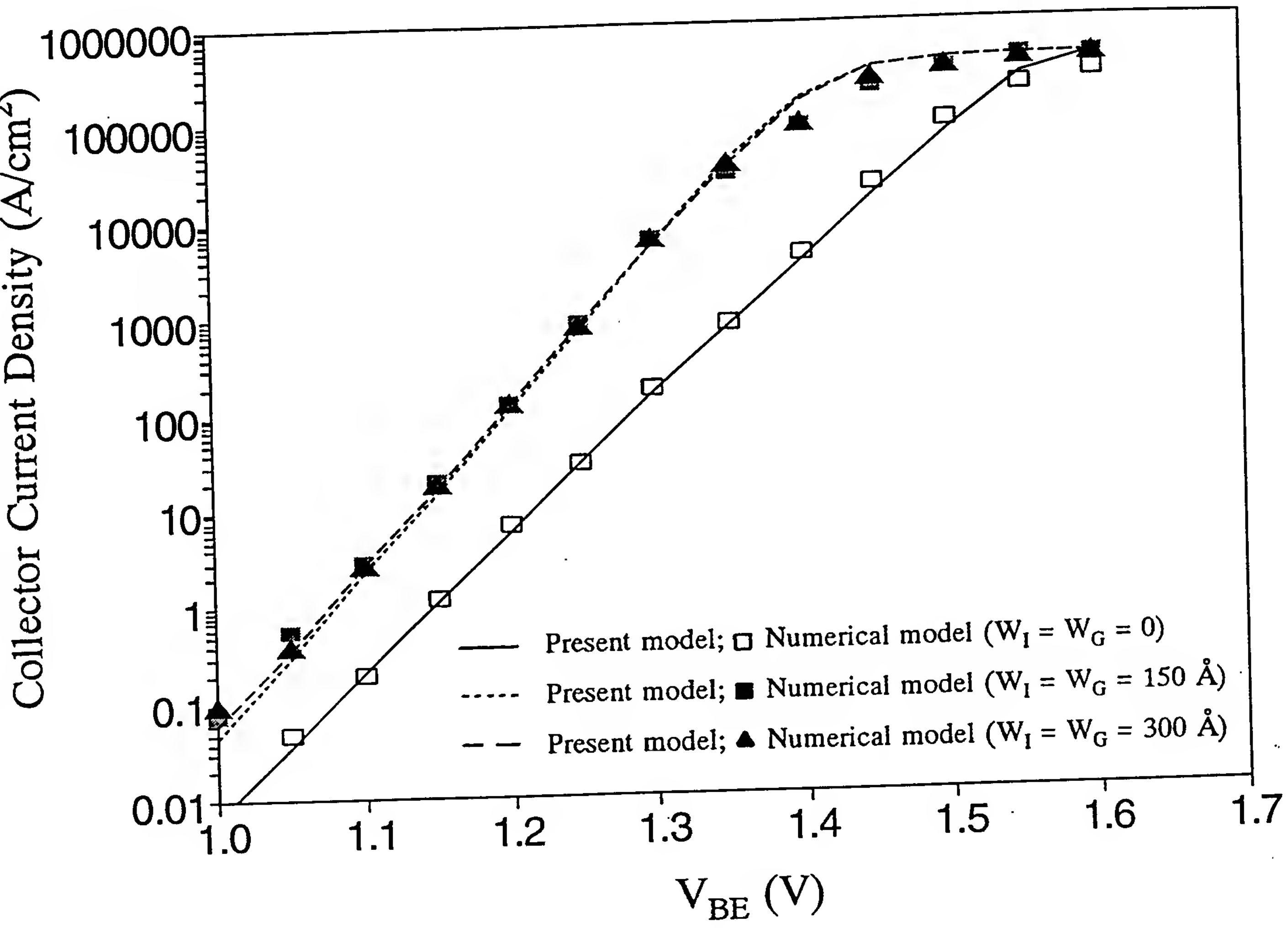


Fig. 4 Collector current densities calculated from the present model and numerical model for HBTs with three different $W_I$ and $W_G$ makeups.

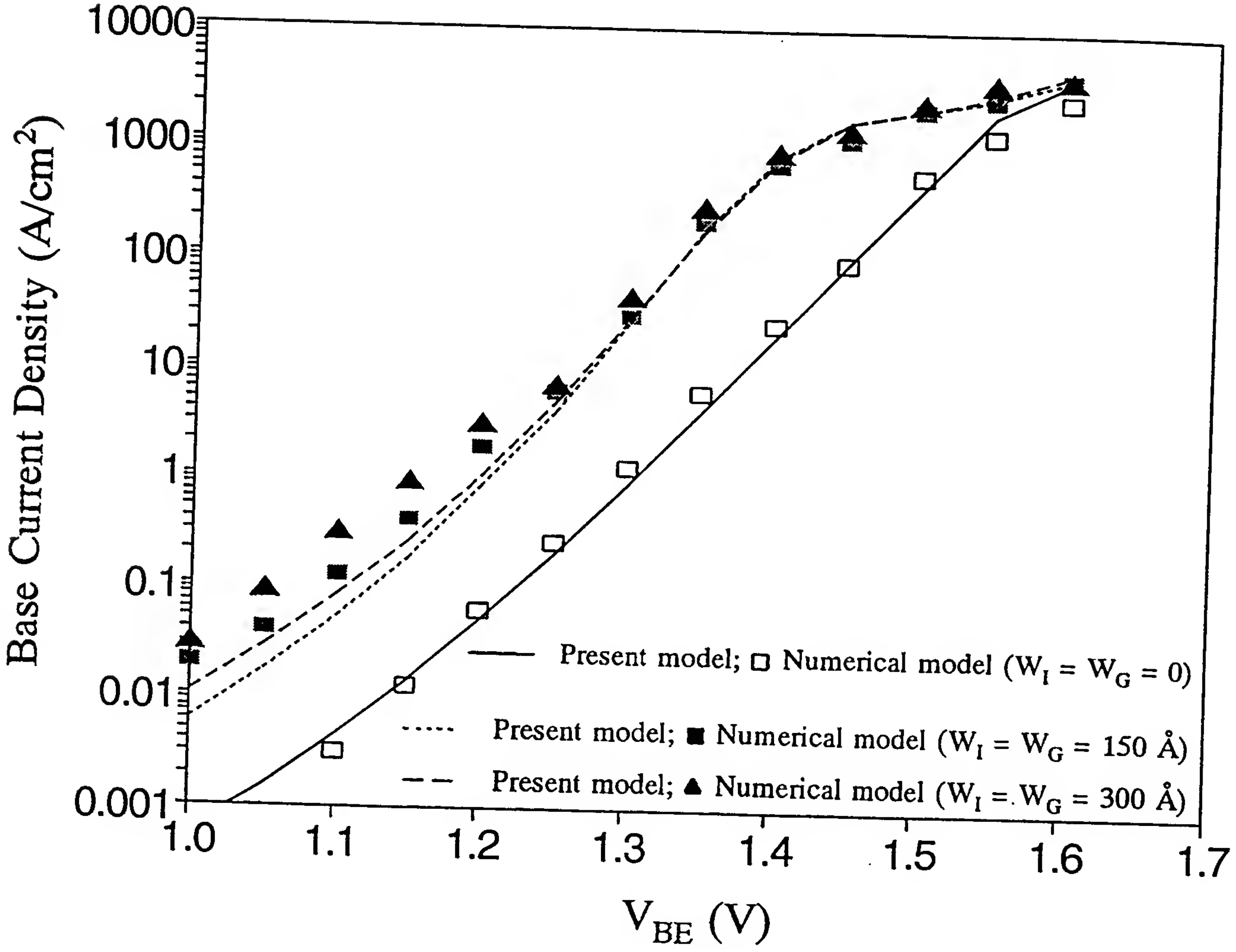


Fig. 5 Base current densities calculated from the present model and numerical model for HBTs with three different $W_I$ and $W_G$ makeups.

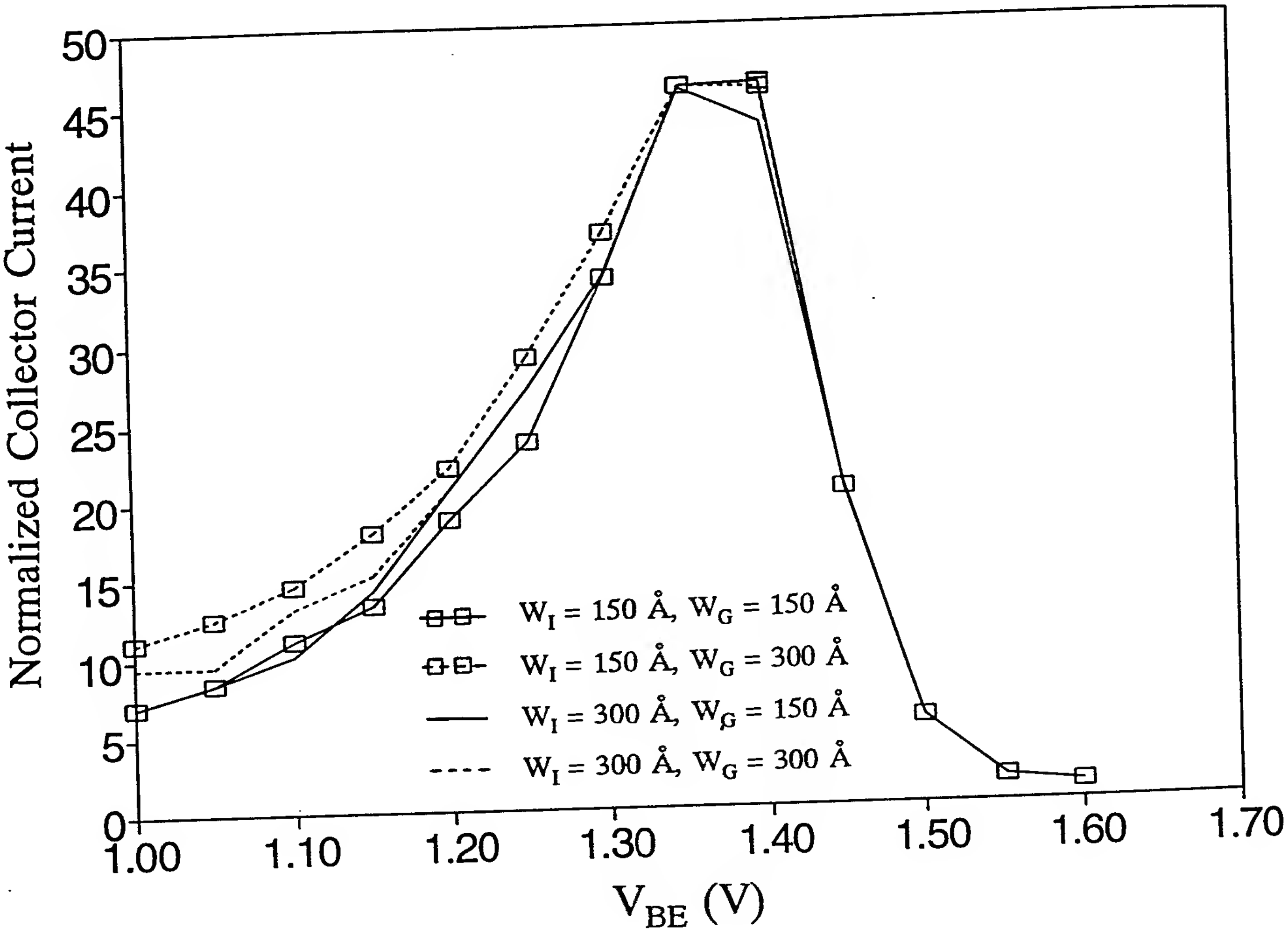


Fig. 6 Collector current densities of HBTs having four different $W_I$ and $W_G$ normalized by that of the abrupt HBT ($W_I = W_G = 0$).

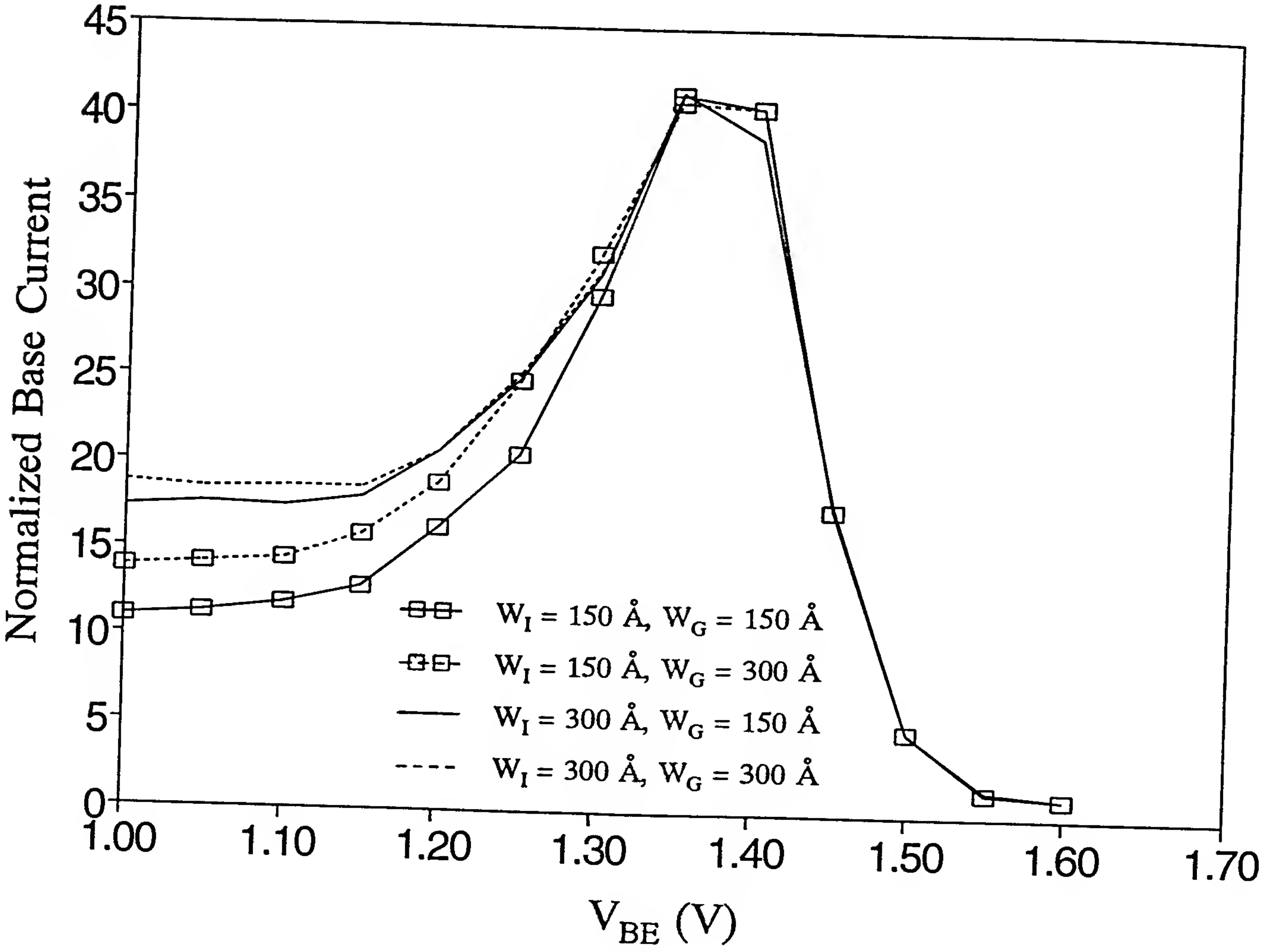


Fig. 7 Base current densities of HBTs having four different $W_I$ and $W_G$ normalized by that of the abrupt HBT ($W_I = W_G = 0$).

high $V_{BE}$. In comparison with the abrupt HBT, $J_C$ in the HBTs with $W_I$ and $W_G$ is about 5-10 times larger at $V_{BE}$ = 1 V, increases to a maximum of about 45 times larger at $V_{BE}$ = 1.4 V, and becomes comparable with that in the abrupt HBT at high voltages. The presence of $W_I$ and $W_G$ also increases $J_B$, as shown in Fig. 7. Intuitively, one would expect that $W_I = W_G$ = 300 Å results in the widest space-charge layer and thus the highest $J_B$. This is indeed the case in Fig. 7, which shows that the HBT with $W_I = W_G$ = 300 Å has the highest $J_B$ at relatively small $V_{BE}$ among the four HBTs. At high $V_{BE}$, all four HBTs have similar $J_B$. This is because the space-charge region has been greatly reduced by the high $V_{BE}$, and the space-charge region recombination is not important. Conversely, the hole current $J_{RE}$ injected from the base to emitter becomes the dominant component for $J_B$. Since $J_{RE}$ is not affected by $W_I$ and $W_G$ at high voltages (the valence barrier is almost flat at such bias conditions), all four HBTs have comparable $J_B$.

Fig. 8 shows the dc current gains at a low current level ($J_C$ = 0.01 A/cm$^2$ is considered) calculated as a function of $W_I$ and $W_G$. The results suggest that the current gain β decreases linearly with increasing $W_G$ if $W_I$ = 0, and β is relatively insensitive to $W_G$ if $W_I$ is greater than 150 Å. Also, at this current level, β decreases with increasing $W_I$ for all $W_G$.

The trends are quite different for the HBT operated at a high current level ($J_C$ = 10$^5$ A/cm$^2$ is considered), however, as shown in Fig. 9. First, comparable current gains are found for all $W_I$ and $W_G$ > 150 Å. Furthermore, when $W_G$ approaches zero, having a non-zero $W_I$ is beneficial, but increasing $W_I$ beyond 150 Å does not provide any current gain enhancement.

## 4. CONCLUSIONS

The spike at the hetero-interface often limits the free-carrier injection from the emitter to base, and the graded and setback layers are frequently used to overcome such a problem. In this study, the combined effects of the graded and setback layers on the current-voltage characteristics of AlGaAs/GaAs HBTs are investigated, and an analytical model which accounts for such effects is developed.

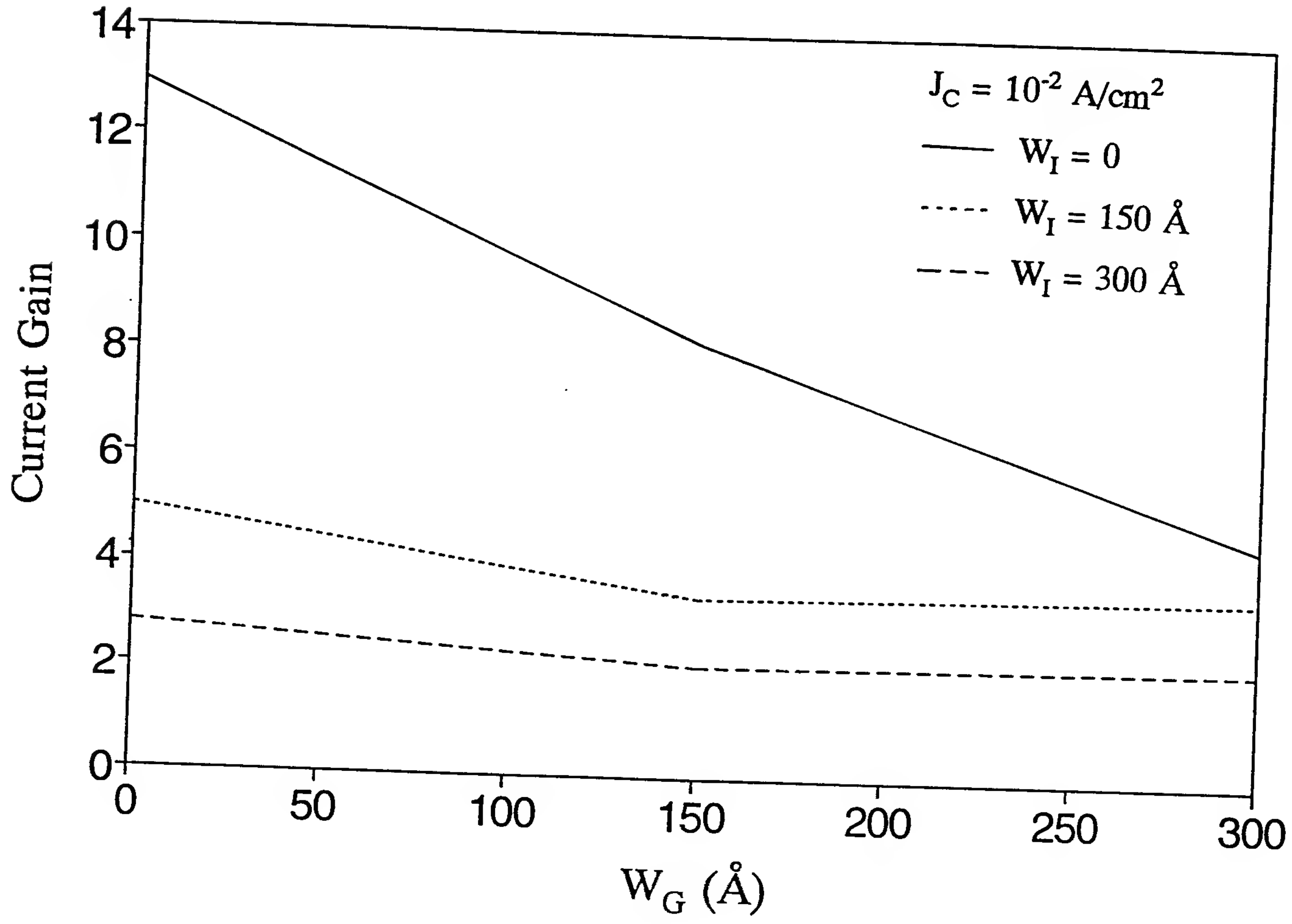


Fig. 8 HBT current gains at a low current level calculated as a function of $W_G$ and $W_I$.

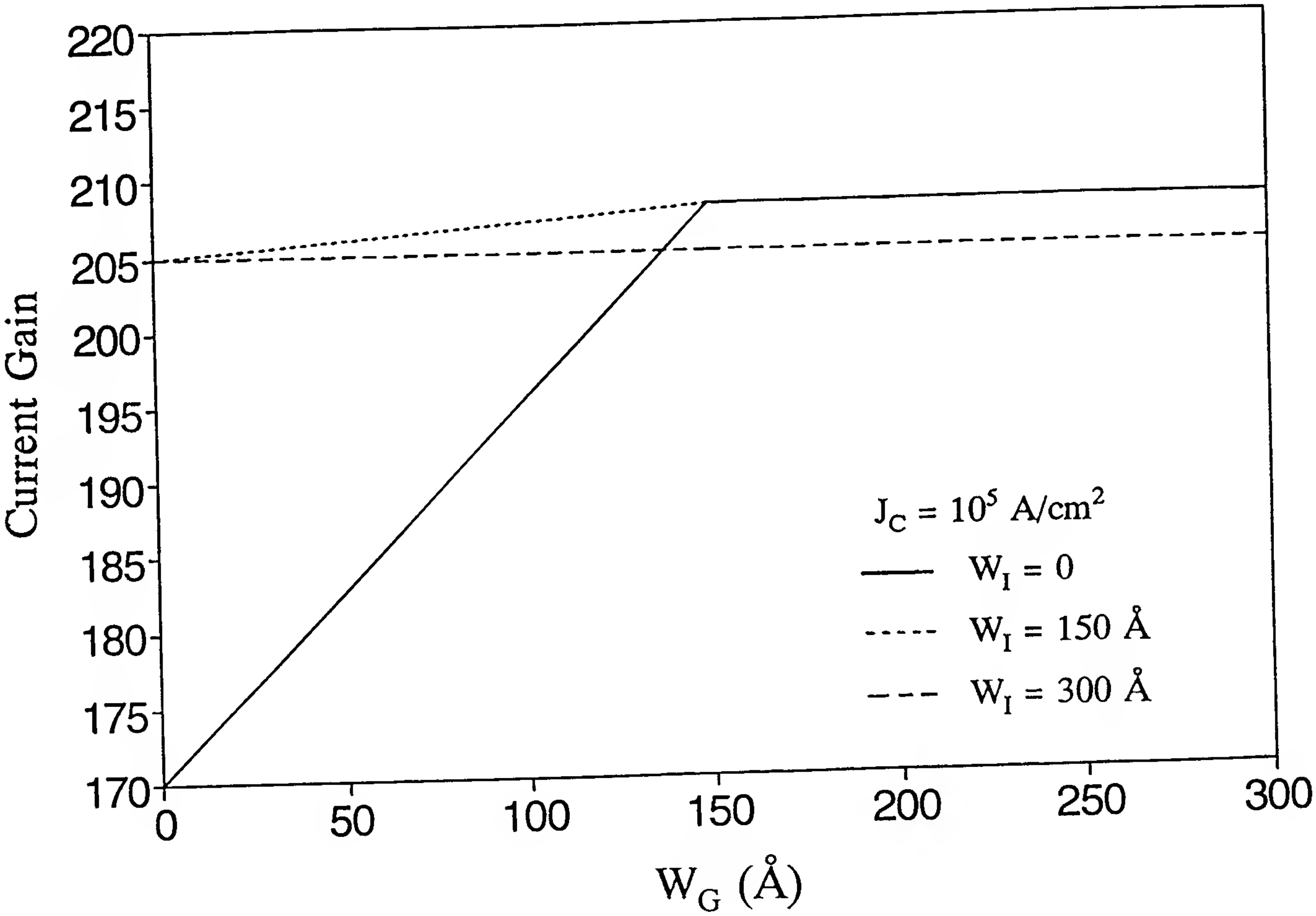


Fig. 9 HBT current gains at a high current level calculated as a function of $W_G$ and $W_I$.

The following conclusions can be drawn from the present study.

1. At low current levels, the presence of $W_I$ and $W_G$ actually degrades the current gain.

2. At high current levels, using $W_I$ = 150 Å and any $W_G$ or $W_I$ = 0 and 150 Å ≤ $W_G$ ≤ 300 Å yields the optimal current gain.

3. Among the HBTs considered (with or without $W_I$ and/or $W_G$), the abrupt HBT ($W_I$ = $W_G$ = 0) possesses the best current gain at low current levels and the worst current gain at high current levels.

The model developed permits a degree of insight into the influence of the graded and setback layers on the HBT performance and should have practical applications for use in HBT design and circuit simulation.

## REFERENCES

[1] H. Kroemer, "Theory of a wide-gap emitter for transistors," Proc. IRE, vol. 45, pp. 1535, 1957.

[2] P. M. Asbeck et al., "Heterojunction bipolar transistors for ultra high speed digital and analog applications," in IEDM Tech. Dig., 1988.

[3] A. Marty, G. E. Rey, and J. P. Bailbe, "Electrical behavior of an Npn GaAlAs/GaAs heterojunction transistor," Solid-St. Electron., vol. 22, pp. 549, 1979.

[4] A. A. Grinberg and S. Luryi, "On the thermionic-diffusion theory of minority transport in heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 40, pp. 859, 1993.

[5] B. R. Ryum and I. M. Abdel-Motaleb, "A Gummel-Poon model for abrupt and graded heterojunction bipolar transistors," Solid-St. Electron., vol. 33, pp. 869, 1990.

[6] J. J. Liou, C. S. Ho, L. L. Liou, and C. I. Huang, "An analytical model for current transport in AlGaAs/GaAs abrupt HBTs with a setback layer," Solid-St. Electron., vol. 36, pp. 819, 1993.

[7] M. S. Lundstrom, "Boundary conditions for pn heterojunctions," Solid-St. Electron., vol. 27, pp. 491, 1984.

[8] A. A. Grinberg, M. S. Shur, R. J. Fischer, and H. Morkoc, "An investigation of the effect of graded layers and tunneling on the performance of AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. ED-31, pp. 1758, 1984.

[9] C. D. Parikh and F. A. Lindholm, "Space-charge region recombination in heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 39, pp. 2197, 1992.

[10] A. Das and M. S. Lundstrom, "Numerical study of emitter-base junction design for AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 35, pp. 863, 1988.

[11] S. -C. Chen, Y. -K. Su, and C. -Z. Lee, "A study of current transport on

p-N heterojunctions," Solid-St. Electron., vol. 35, pp. 1311, 1982.

[12] L. L. Liou and C. I. Huang, "Using constant base current as a boundary condition for one-dimensional AlGaAs/GaAs heterojunction bipolar transistor simulation," Electron. Lett., vol. 26, pp. 1501, 1990.

[13] A. Chatterjee and A. H. Marshak, "Theory of abrupt heterojunctions in equilibrium," Solid-St. Electron., vol. 24, pp. 1111, 1981.

[14] D. A. Sunderland and P. L. Dapkus, "Optimizing N-p-n and P-n-p heterojunction bipolar transistors for speed," IEEE Trans. Electron Devices, vol. ED-34, pp. 367, 1987.

[15] W. Liu and J. S. Harris, Jr., "Diode ideality factor for surface recombination current in AlGaAs/GaAs heterojunction bipolar transistors," IEEE Trans. Electron Devices, vol. 39, pp. 2726, 1992.

[16] S. Tiwari and D. J. Frank, "Analysis of the operation of GaAlAs/GaAs HBT's," IEEE Trans. Electron Devices, vol. 36, pp. 2105, 1989.

# EXPERIMENTAL STUDY OF A SWITCHED RELUCTANCE MOTOR

Shy-Shenq P. Liou
Assistant Professor

Division of Engineering
San Francisco State University
1600 Holloway Avenue
San Francisco, CA 94132

Final Report for: Summer Research Extension Program

The author would like to thank the sponsorship of
Air Force Office of Scientific Research
Bolling AFB, Washington DC
and
San Francisco State University

June, 1994

# EXPERIMENTAL STUDY OF A SWITCHED RELUCTANCE MOTOR

Shy-Shenq P. Liou
Assistant Professor
Division of Engineering
San Francisco State University

## Abstract

Several search coils, four for phase windings and one for yoke, were installed into a prototype switched reluctance motor to identify the inductance of phase windings and yoke search coil with respect to the relative position of rotor and stator teeth. The induced voltage waveforms for these search coils when the tested switched reluctance motor is running at constant speed mode are also measured and reported. Three temperature sensors were also installed with the intention to measure the temperature profile of the switched reluctance motor. Due to the malfunction of these temperature sensors, no temperature profile is reported. A four phase, switched reluctance motor driver using MCTs are also designed and built. The driver circuit for the MCT is also reported.

# EXPERIMENTAL STUDY OF A SWITCHED RELUCTANCE MOTOR

Shy-Shenq P. Liou

## INTRODUCTION

Switched reluctance motor is a relative new member of the family of rotating machines. Although conceptually the switched reluctance motor is a simple electromechanical energy conversion device, the popularity of it nowadays is not possible without the pioneering work of Professor P.J. Lawrenson of University of Leeds [1] almost 20 years ago and the advance of semiconductor switching devices.

Basically, switched reluctance motor is an electric motor in which torque is produced by the tendency of its movable part, the rotor, to move to a position where the inductance of the excited winding is maximized. In energy conversion terminology, torque is produced because of its tendency to maximize the coenergy [2].

Figure 1 shows a typical switched reluctance motor in a 8-6 design in which there are eight stator teeth and 6 rotor teeth. There are also four phase windings wrapped around those eight stator teeth; one winding for each two stator teeth. Only one phase winding with one turn each on the stator teeth is shown in Figure 1. Also shown in the Figure 1 are two positions, aligned and unaligned positions. It can be expected that the inductance for the phase winding shown in Figure 1 will assume a maximum value whereas the second phase winding to the right of the shown winding will assume a minimum inductance. If a DC current is supplied to the first phase winding to the right of the winding shown in Figure 1, the switched reluctance motor's rotor will rotate counterclockwise because of the tendency of maximizing the inductance of the energized phase winding. Usually, only one phase winding is energized at any given time. Also it should be noted that only DC current is needed to produce torque.

To produce a continuous rotation, current should be supplied to successive phase winding one at a time based on the position of the rotor with respect to the stator. Therefore, some kind of switching mechanism must be there to TURN ON and TURN OFF the current to each phase winding with reasonable efficiency and power handling capability. The advance made by the semiconductor industry on power switching devices make the control of switched reluctance motor possible. Also, it is important to note that there is no need to put any coil on the rotor poles to create the energy conversion. This turns out to be another salient feature of switched reluctance motor.

Two typical power electronic phase switching circuits for switched reluctance motor are shown in Figure 2. Mos-Controlled Thyristor (MCT) are used in this instance just for representation purpose only. Usually, either MOSFETs or IGBTs are used in the commercial products. Two switches must be turned on by gating from the control circuit simultaneously in order to establish the current flow in the phase winding. When the switches are turned off, the stored energy in the phase winding is returned to the DC source via the two freewheel diodes as shown in Figure 2. The control circuitry must get the rotor position through either encoder or resolver before it can decide which phase winding to be turned on.

The uniqueness of switched reluctance motor can be summarized as the following:

1. There is no winding on the rotor poles.

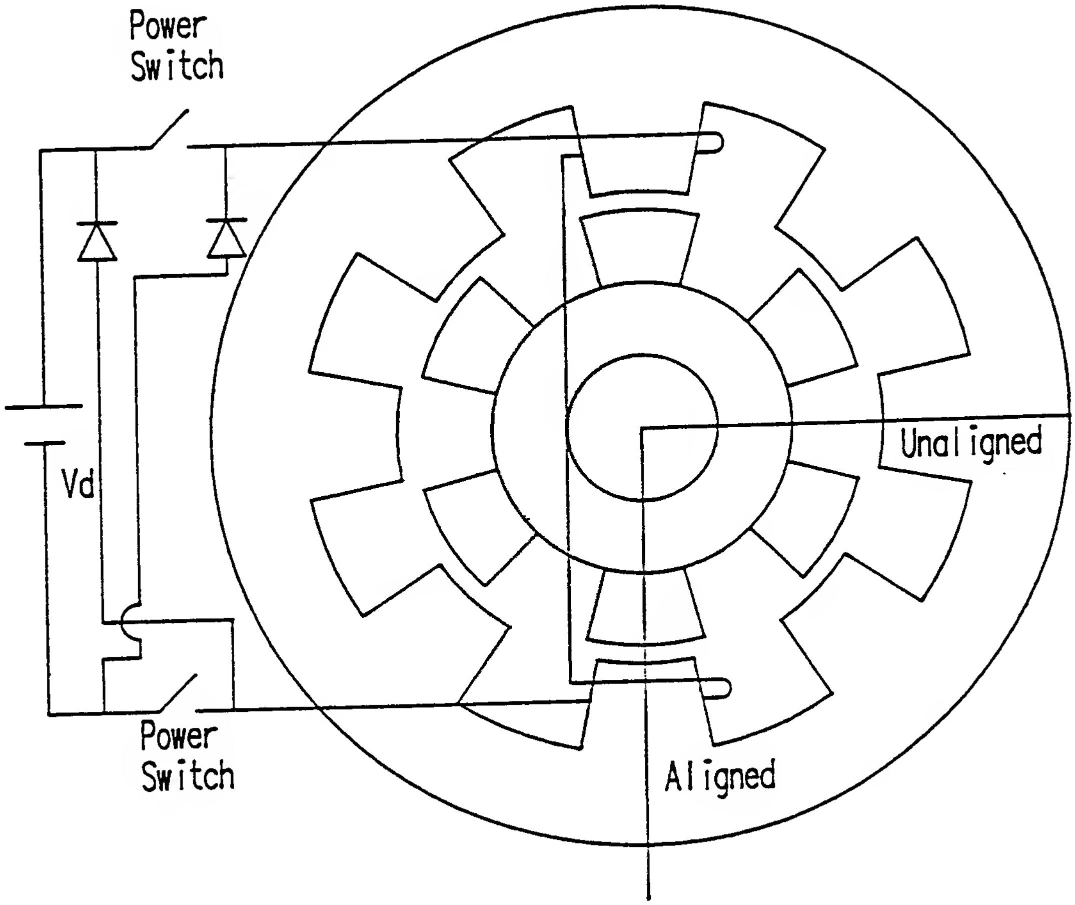


Figure 1: An Eight-Six Design Switched Reluctance Motor

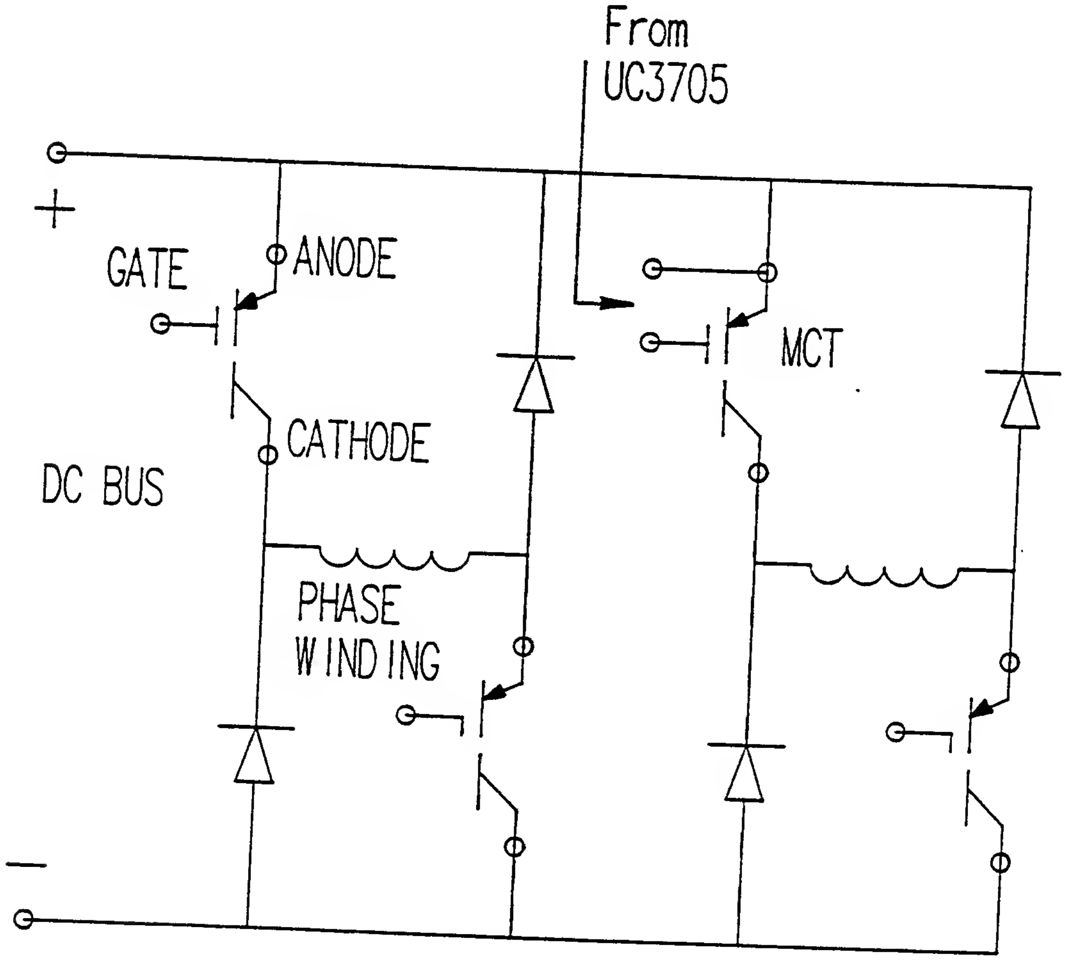


Figure 2: Typical Power Electronic Circuit for Switched Reluctance Motor

2. Both stator and rotor have salient poles.
3. The stator winding comprises a set of coils, each of which is wound on the pole.
4. Excitation is a sequence of current pulses applied to each phase in turn.
5. As the rotor rotates, the phase flux linkage should have a triangular or sawtooth waveform [3] but not vary with current.

## POTENTIAL PROBLEMS with SWITCHED RELUCTANCE MOTORS

The switched reluctance motor has many advantages such as simple geometry, easy to manufacture, high temperature capability, and inherent fault tolerant operation, etc. On the other hand, it does have many undesired features too. First of all, it requires a rotor position sensor (either encoder or resolver). Second, it definitely requires a dedicated power electronic controller. Last but not least, the simple yet complicated magnetic geometry poses a serious engineering challenge for design engineer to design and optimize a switched reluctance motor for specific application. Therefore, the potential problems associated with the switched reluctance motor are summarized in the following:

1. Needs a rotor position sensor in the form of either encoder or resolver. (No dedicated sensor operation should be ideal).
2. How to model the switched reluctance motor electrically and magnetically due to complicated magnetic geometry and severe local saturation when rotor teeth approach those of stator.
3. How to model the switched reluctance motor thermally so design engineer can optimize the

design for given specific application.

Some of the potential problems mentioned above will be discussed in this report. Attempts were made to address as many problems as possible. Due to equipment limitation, not every problem area is attacked during the course of this research. But the effort lay a good foundation for future research work.

## PROTOTYPE SWITCHED RELUCTANCE MOTOR

A prototype switched reluctance motor made by Magna Physics in Hillsboro, Ohio, Model No. SR-90 was purchased and used in all the tests. The specification for this switched reluctance motor is

| | |
|---|---|
| No. of Phases | 4 |
| Voltage | 160 volts DC |
| Peak Output | 400 Watts |
| Rated Speed | 3,000 RPM |
| Maximum Speed | 15,000 RPM |
| Rated Torque | 100 Oz-in |
| Max. RMS Current | 3.0 Amps |
| Efficiency | 80% |
| Encoder | HP HEDS-5000 |
| Resolution | 360 Lines/Revolution |
| Weight | 5.5 Lbs |

A controller from Semifusion Company, Santa Clara, CA was also purchased to control the switched reluctance motor. It allows the user to do various operating mode, position, velocity, torque, and profile modes. The input is menu-driven and very user friendly. A IBM compatible personal computer is required for user interface between the user and controller. A RS232 I/O protocol is used in this regard. A 40-pin Hewlett Packard HCTL-1100 general purpose motion control chip is used in the controller to take care of the number crunching task. For example, read the rotor position from the position encoder and output the proper commutating sequence to individual phase winding based on the desired command operating mode. This setup turns out to be extremely useful for initial testing of the switched reluctance motor.

A simple friction type dynamometer from Hampton Inc. is used to load the switched reluctance motor during the tests. There is no constant torque load from this dynamometer. Therefore, only constant speed command operating mode is reported. In the future, if a decent dynamometer is available, more tests on the constant torque operating mode can be tested.

## INDUCTANCE MEASUREMENT THROUGH SEARCH COILS

Five search coils were put inside of the switched reluctance motor in order to identify various parameters associated with the motor. One aim is to see whether the inductances of these search coils change when the rotor position varies. If there is a unique pattern or certain fingerprint associated with the rotor position, this might provide an easy way to locate the position of the rotor without either the encoder or resolver. Four one-turn search coils were put into the same locations with the phase windings. Another search coil enclose the yoke or back iron portion of the motor. A handhold digital inductance meter was used to measure the inductances of these search coils.

Because the air gap of the switched reluctance motor is small and localized saturation is very severe, it is important to fix the relative position between stator and rotor while inductance measurement was conducted to avoid any inaccuracy.

Attempt was made to run the switched reluctance motor controller in position operating mode to fix the location of the rotor. There are two problems with this approach. First, the resolution is about 2.5 degrees even although user can specify a resolution of about 0.25 degree in the menu-driven PC program. Second, the DC current flowing in one of the phase windings provide a DC offset for the magnetic field inside the switched reluctance motor. This makes the inductance meter reading extremely unstable and hence render it useless.

A rotary table with a resolution of minimum one degree is set up as shown in Figure 3 to measure the inductances of search coils more accurately. A C-clamp is used to fix the rotor position at standstill whereas the stator part is adhered to the rotary table. For every degree adjustment from the rotary table, measurements were made for each search coils. The rotary table has a 96 degrees adjustable range. For a four phase, 8-6 design switched reluctance motor, minimum range for the angle is 360/6=60 degrees. Thus it provides enough data points for a complete cycle. The inductances of each search coil versus the rotor angle for phase winding one to four and for yoke search coil are shown in Figures 4 to 8 respectively.

From Figures 4 to 7, it can be seen that the inductances of phase winding search coils do vary cyclically with the rotor position with a distinct difference when the rotor teeth is aligned with the stator teeth (minimum inductance); when the rotor is at aligned position (maximum inductance); and in the transition stage between the maximum and minimum inductance. Also it repeats roughly every 60 mechanical degrees. Based on the shape of these inductance curves, it seems promising that maybe there is some uniqueness in these

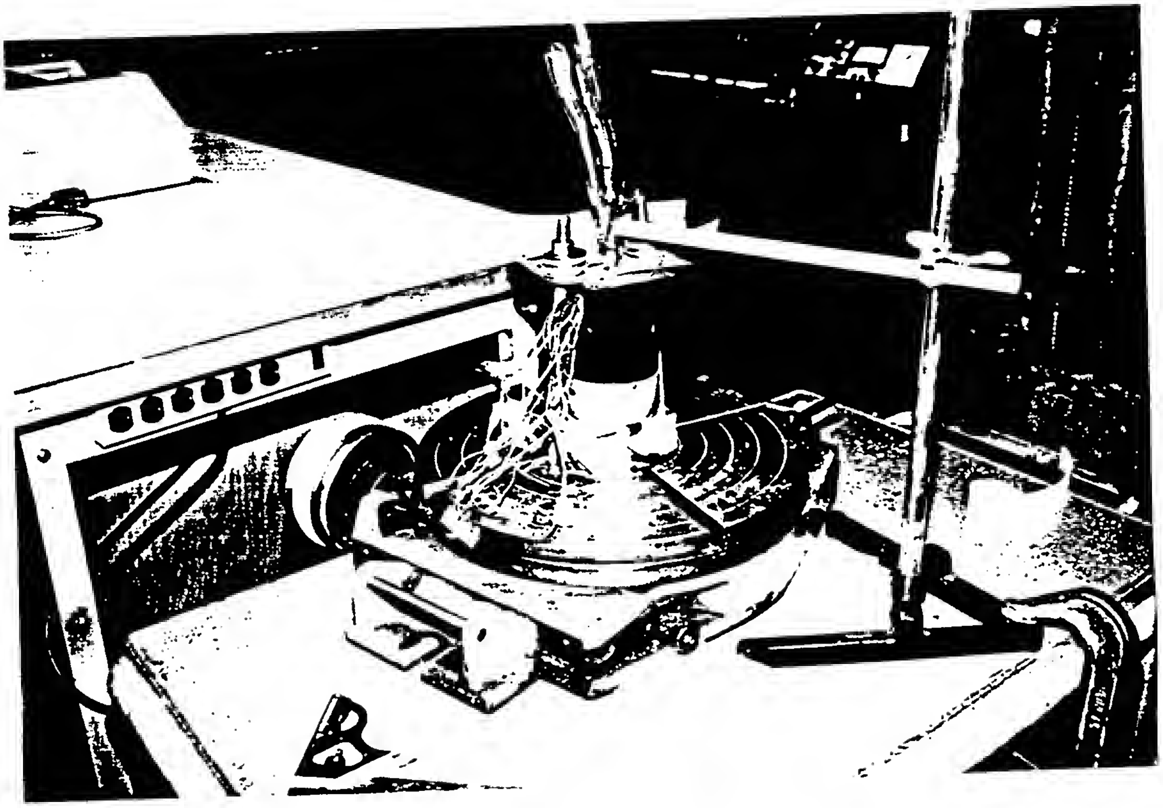


Figure 3: Rotary Table Setup for Inductance Measurement

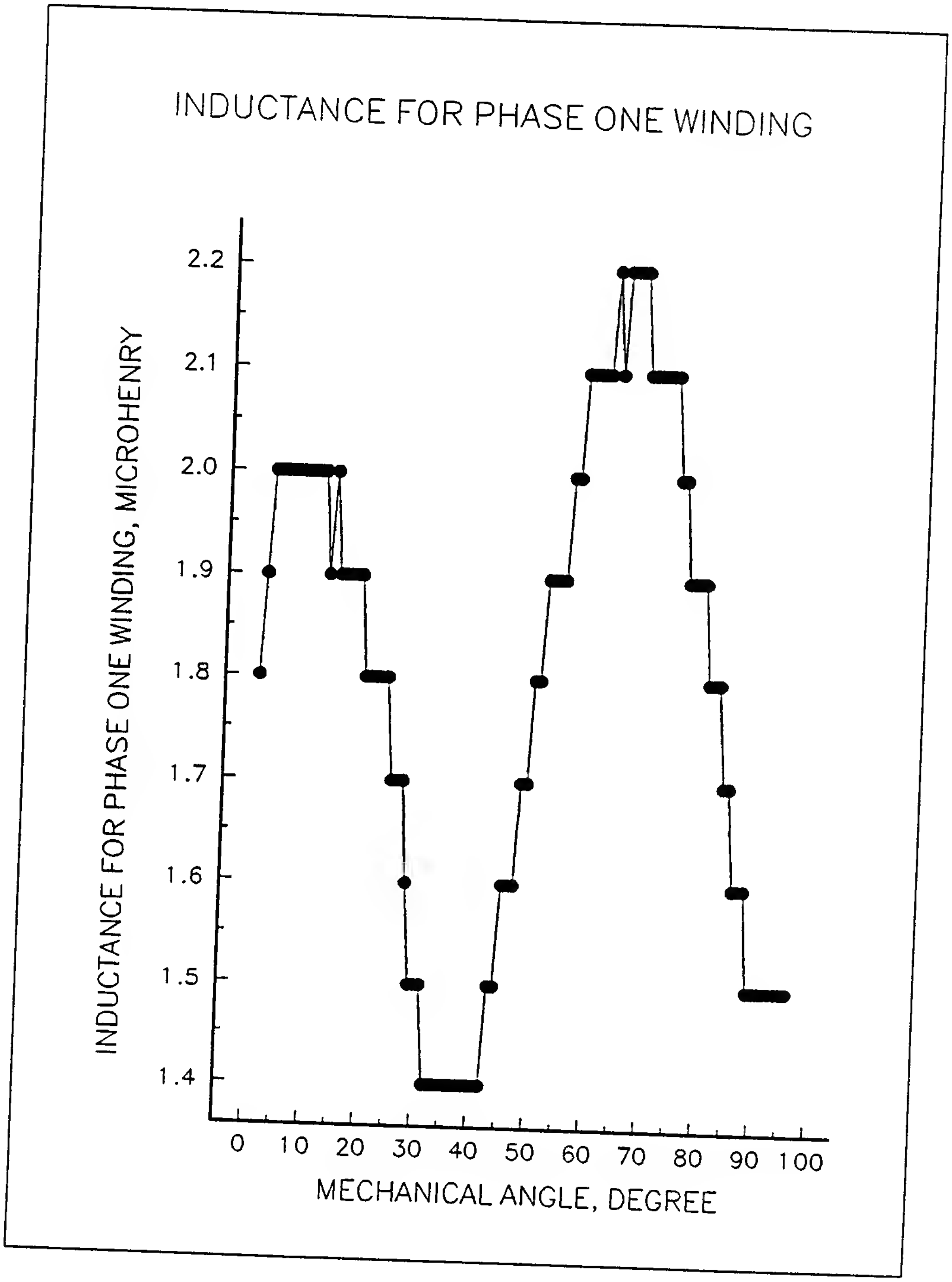


**Figure 4: Inductance versus Rotor Angle for Phase One Search coil**

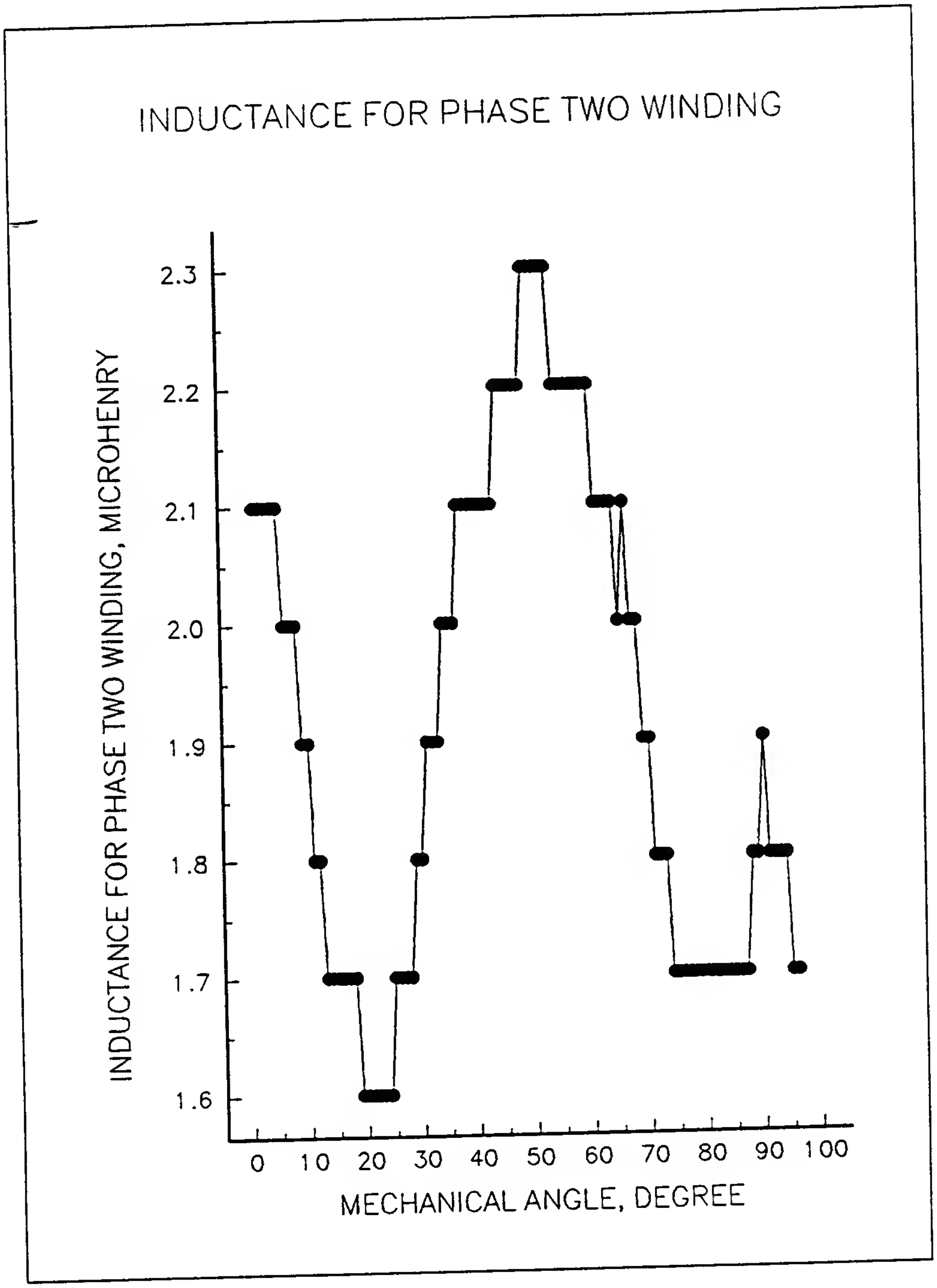


**Figure 5: Inductance versus Rotor Angle for Phase Two Search coil**

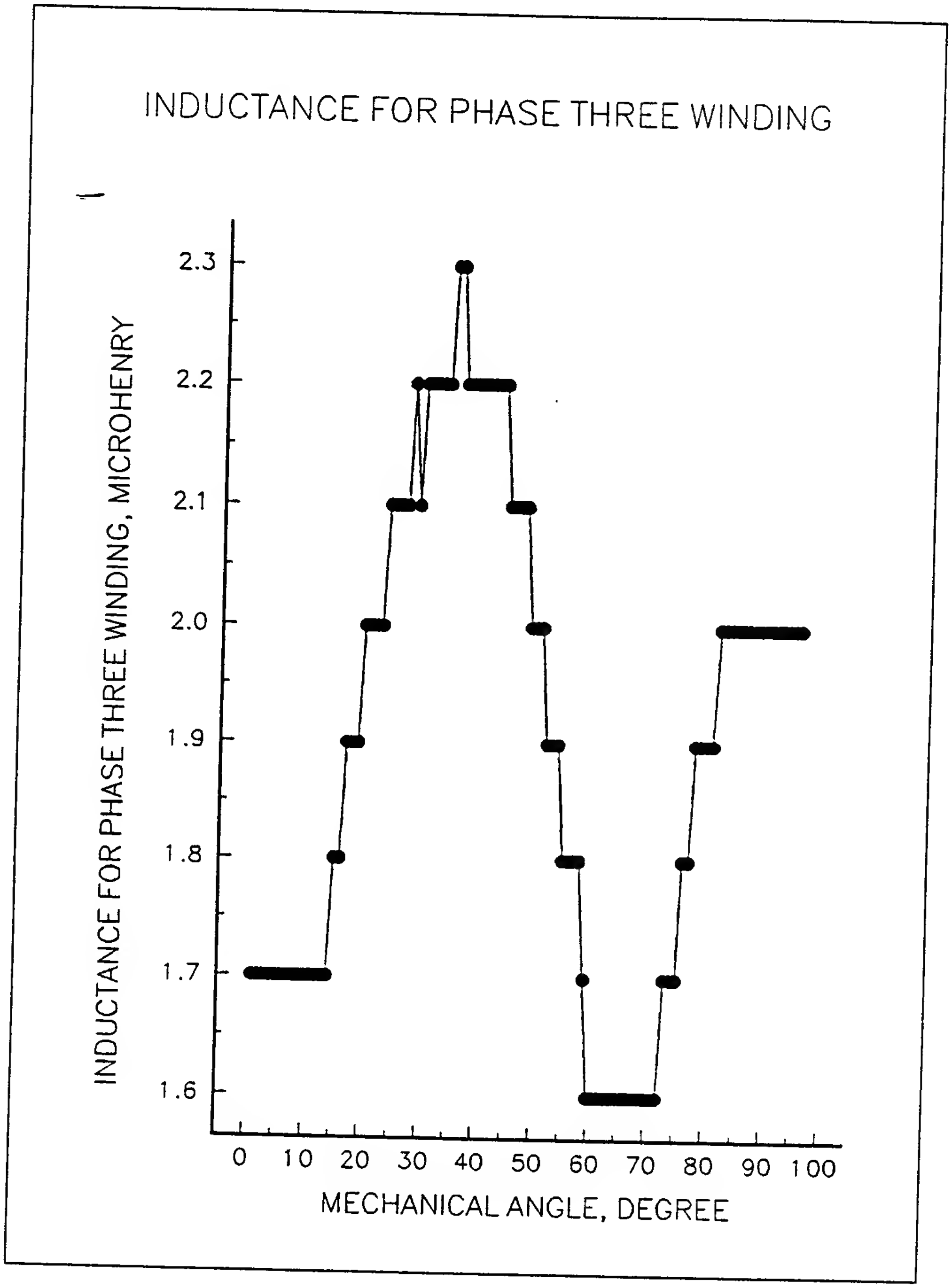


**Figure 6: Inductance versus Rotor Angle for Phase Three Search coil**

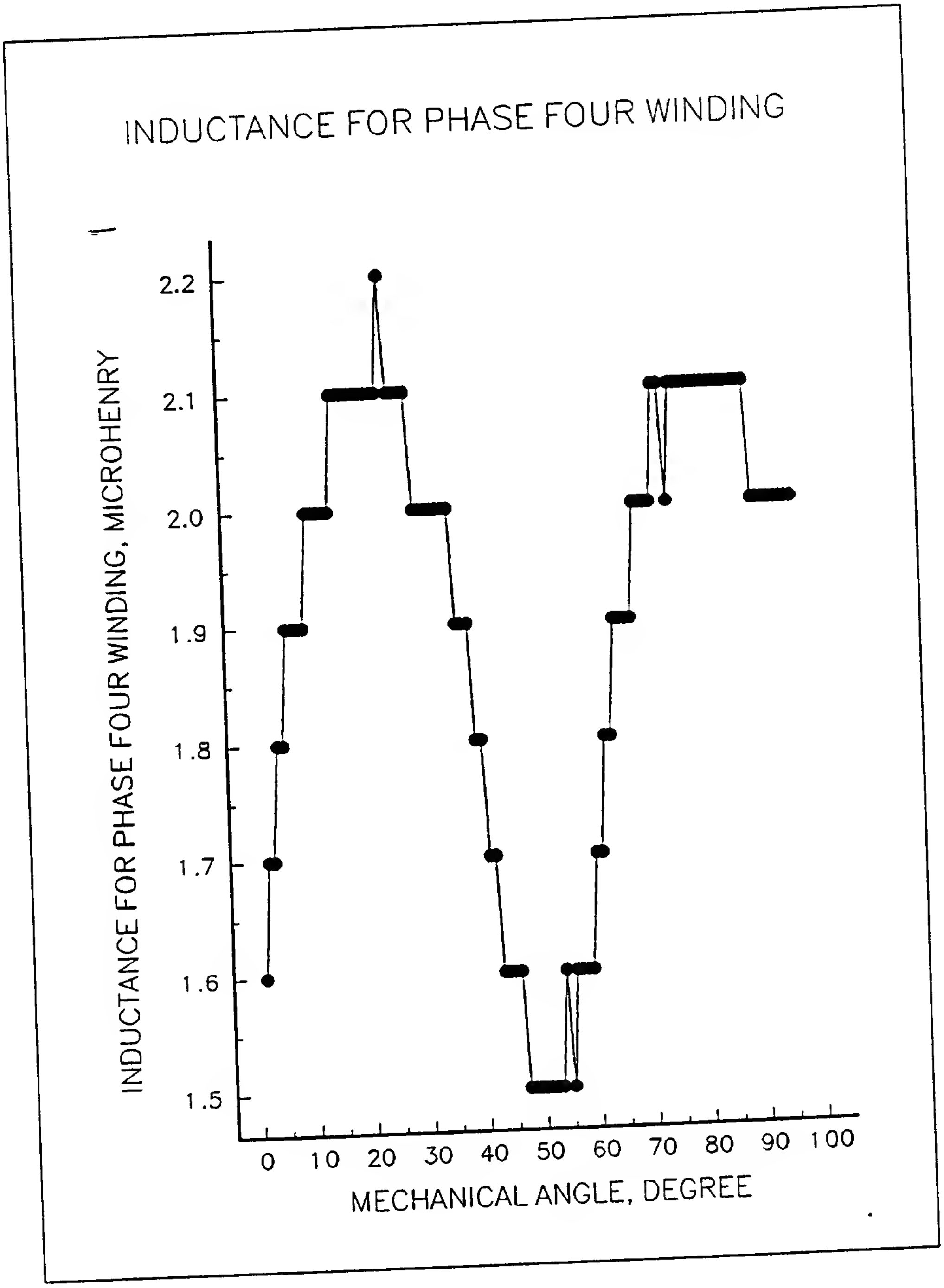


**Figure 7: Inductance versus Rotor Angle for Phase Four Search coil**

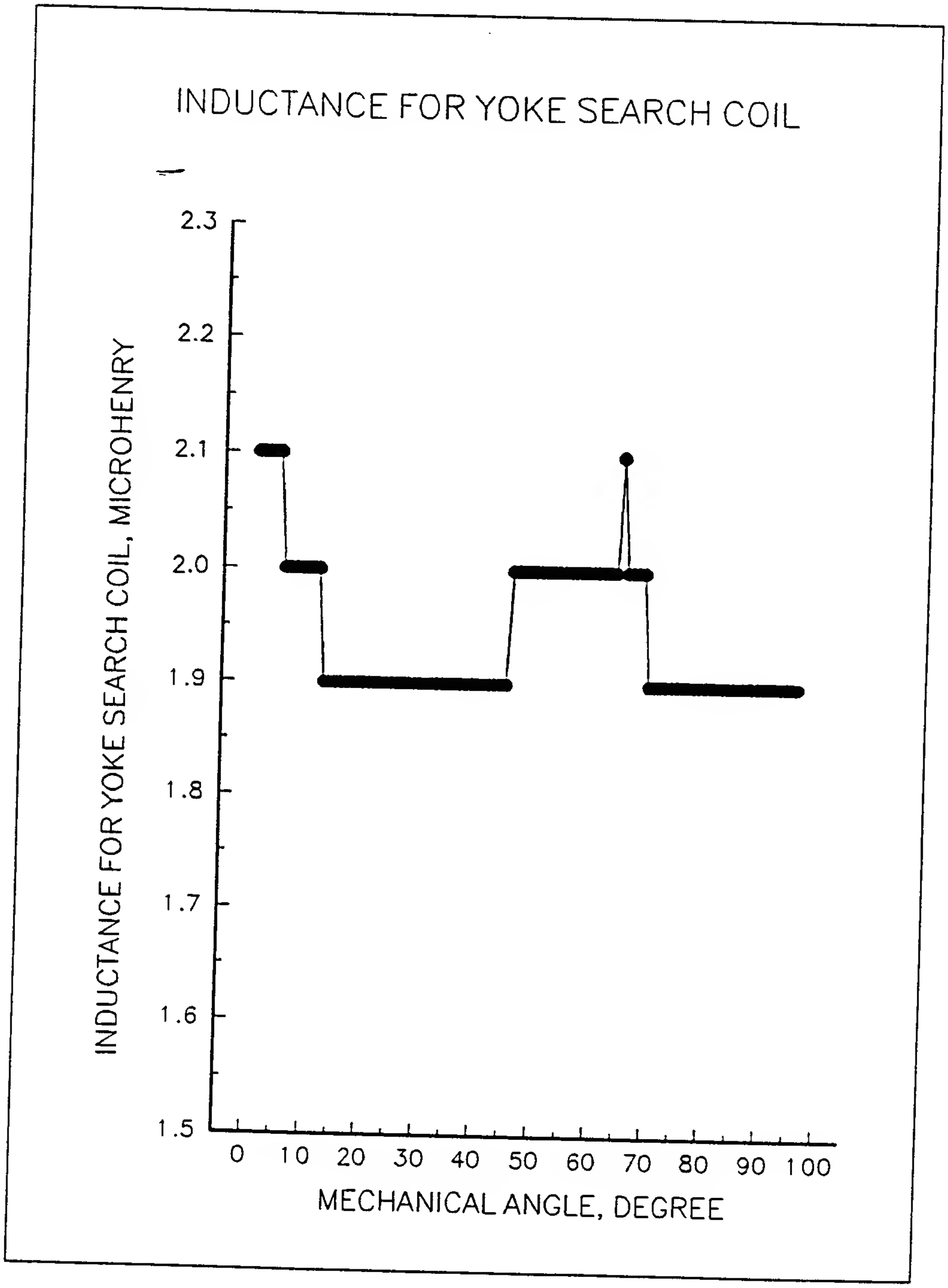


**Figure 8: Inductance versus Rotor Angle for Yoke (Back iron) Search coil**

inductance patterns. It is also worthwhile noticing that the inductance value is quite small, in the range of 1.3 to 2.3 microHenry only. This is because there is only one turn in each search coil. If a dedicated search coils were to be installed inside the motor to serve as the position detector, the number of turns shall not be too big. Thus, the absolute inductance value probably is inherently small to begin with. Consequently, emphasis shall not be to boost the magnitude of the measured inductance through the search coils. Instead, the focus shall be placed on the slope of these inductance values since it provides a distinct and unique characteristic.

If a digital circuitry is designed to filter these inductance values to produce a rotor position by assigning 1 to a rising inductance and 0 to a decreasing inductance, the maximum resolution this can be achieved will be

$$360/(2^N) = 360/(2^4) = 22 \text{ degrees} \qquad (1)$$

where N is the number of search coils.

It is hoped that the inductance measurement of the search coil on the yoke can provide additional insight into the rotor position identification. By examining the Figure 8, it is determined that the inductance of the yoke search coil does not vary sufficient to make any impact on the position sensing at all.

**FLUX WAVEFORMS**

A digital oscilloscope is used to measure the induced voltage waveforms for each search coil for the stator poles and yoke at various operating conditions and RPM. A 2000 RPM at no load was tested first just to observe the waveforms.

Because the oscilloscope can only measure the induced voltage, not the flux linkage for the search coils, all the waveforms shown here are voltage waveforms, not the flux linkage waveforms. A integration operation must be done to get the flux linkage waveform from the induced voltage waveform. This can be achieved by store the data points on the disk and then use some routine to dump the data into a computer then a software can be written to do the numerical integration of the captured induced voltage waveform. This process is currently underway.

Figures 9 and 10 shows the induced voltage waveforms for phase winding 1 and 4 respectively. Same triggering is used to capture these two waveforms. It can be seen that they are exactly the same other than the time delay because of different control triggering signals sent by the HCTL-1100 motion control chip. The time difference is about one minisecond in this instance.

An expanded waveform for Figure 9 is shown in Figure 11 for detailed examination. Keep in mind that the waveform shown is the induced voltage instead of flux linkage waveform. There are three short positive pulses and then a relatively big negative pulse. A constant pulse produces a triangular waveform after the integration process. If the waveform shown in Figure 11 is integrated to yield the flux linkage waveform, the expected flux linkage waveform will be a triangular one as reported in reference [4]. Also, the flux must reset itself somehow, therefore the total enclosed area from the positive pulses and negative pulse would be approximately the same as evident in Figure 11. Once again, there is almost no torque required by the load except the friction and windage losses of

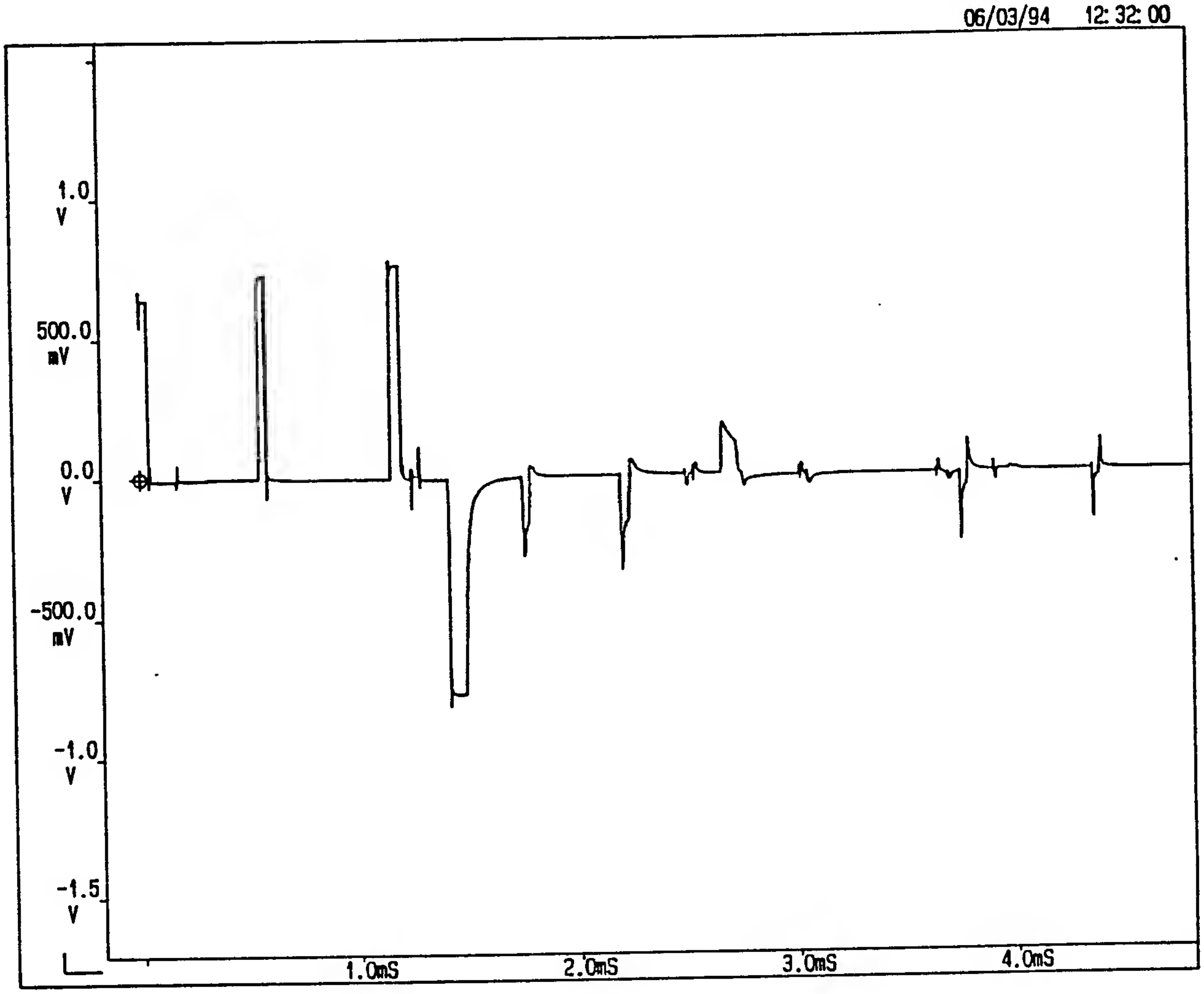


**Figure 9: Induced Voltage Waveform for Phase One Search Coil**

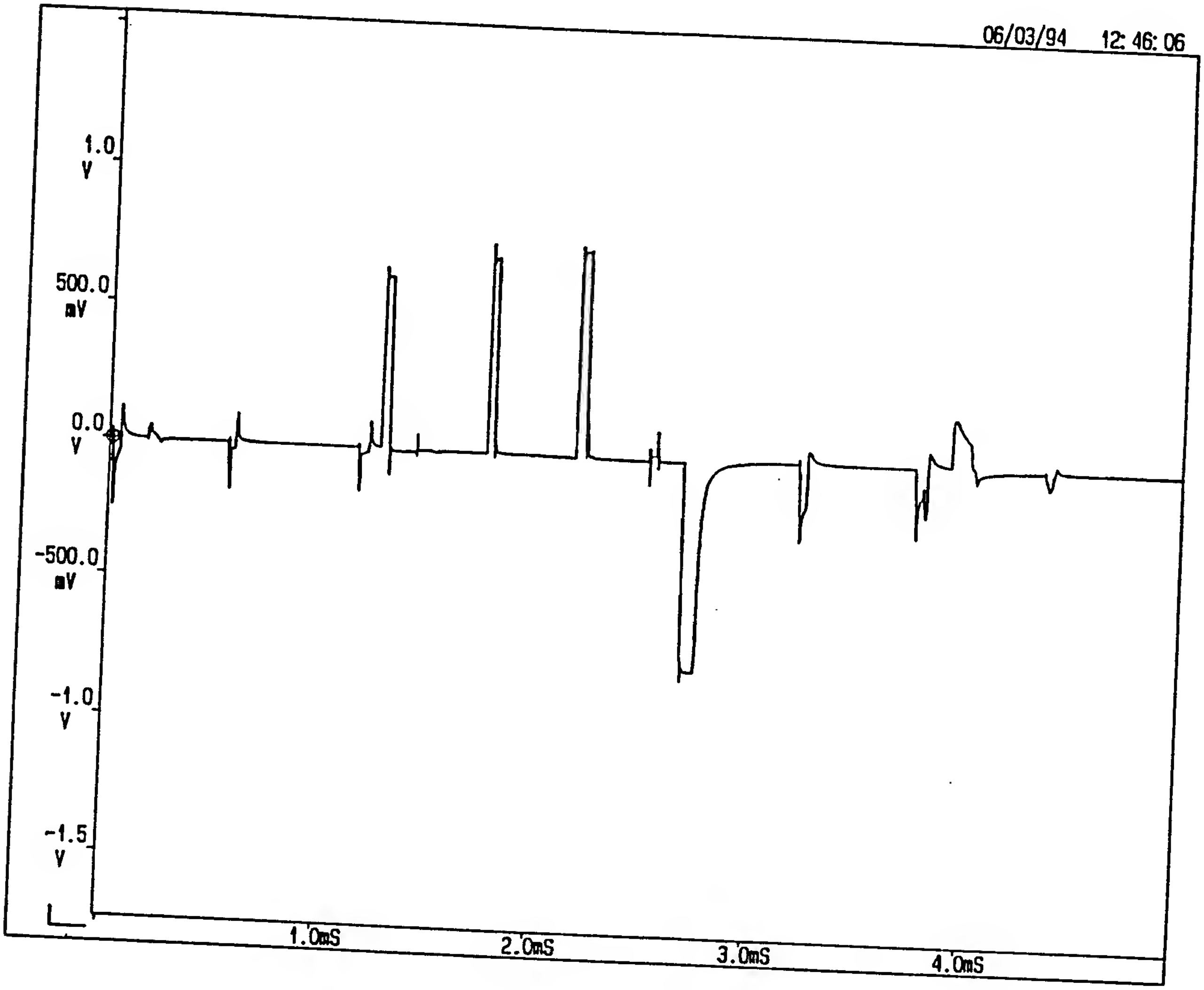


Figure 10: Induced Voltage Waveform for Phase Four Search Coil

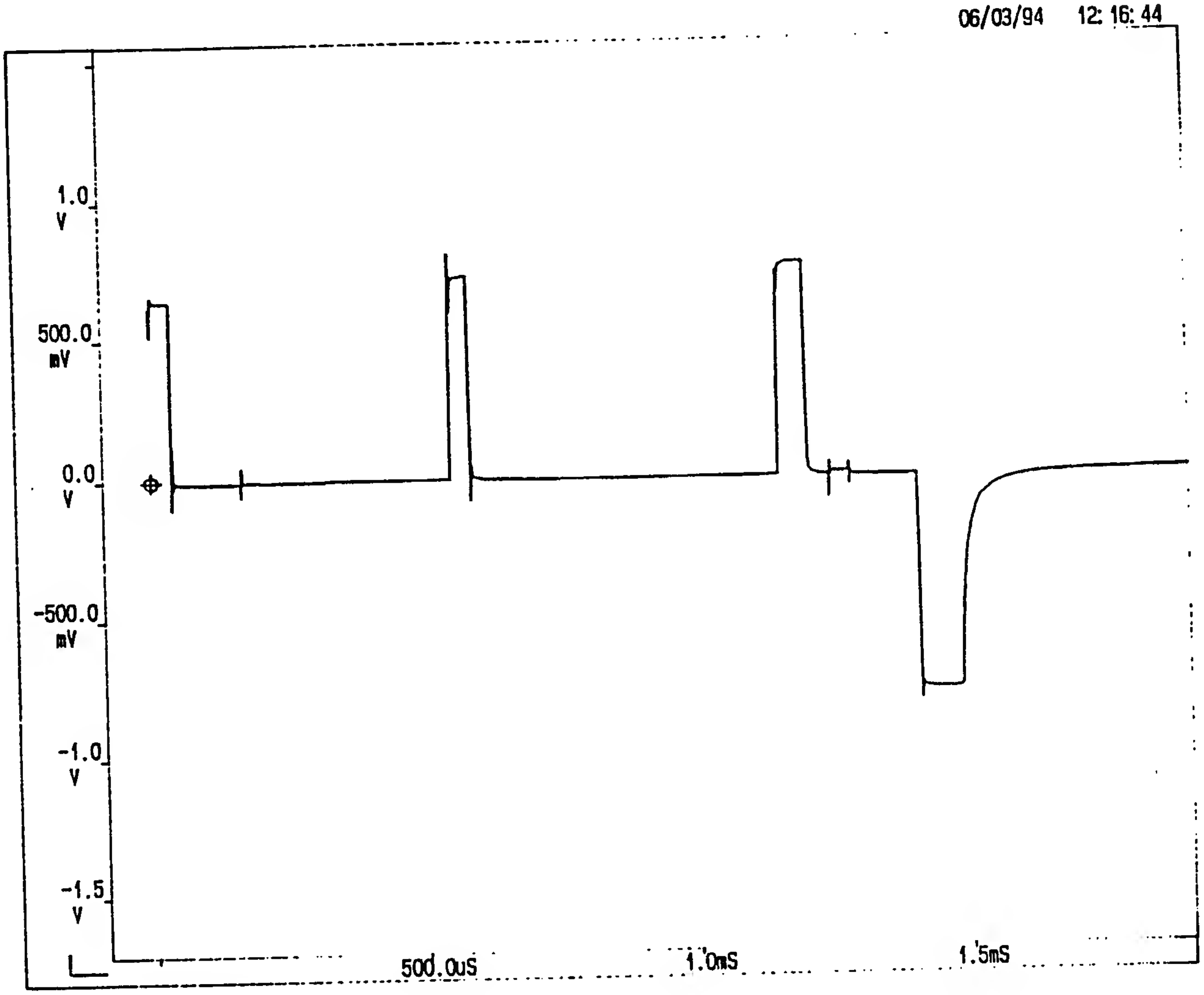


**Figure 11: Expanded Induced Voltage Waveform for Phase One Search Coil**

the motor. Thus, very small current is required to accelerate the motor to the desired speed. Therefore, the flux established in the air gap is relatively small.

The induced voltage waveform from the search coil on the yoke (back iron) is shown in Figure 12. Basically, there is absolutely no distinct pattern for this waveform. It is extremely difficult to imagine how the flux waveform would look like without actually conducting the integration process. One sure thing is that since the induced voltage waveform consists of many constant voltage pulses, the flux linkage waveform will be sawtooth like. It is also quite difficult to identify the complete cycle for this induced voltage waveform. Maybe once the flux linkage waveform is obtained, a more distinguishable pattern can be found.

When the motor is loaded at about 40% of the rated load at about the same speed (2000 RPM), induced voltage waveforms were taken for each search coils and the waveforms for phase one and yoke search coil are shown in Figures 13 and 14 respectively. It can be seen from these two figures that the motor is already operating in the current chopping mode. A series of short positive pulses followed by a relatively long negative pulse finish one firing of phase winding. Also, the waveform is quite noisy. Once the integration is conducted, the waveform will look much nicer.

Once again, no distinct pattern can be found from Figure 14 to identify the cycle of either induced voltage or flux linkage waveform. Based on the observations from Figures 12 and 14, the flux linkage waveform at the yoke (back iron) section of the motor probably is highly dependent on the location of the search coil and what is the operating mode of the motor. It might not be easy to predict the flux linkage

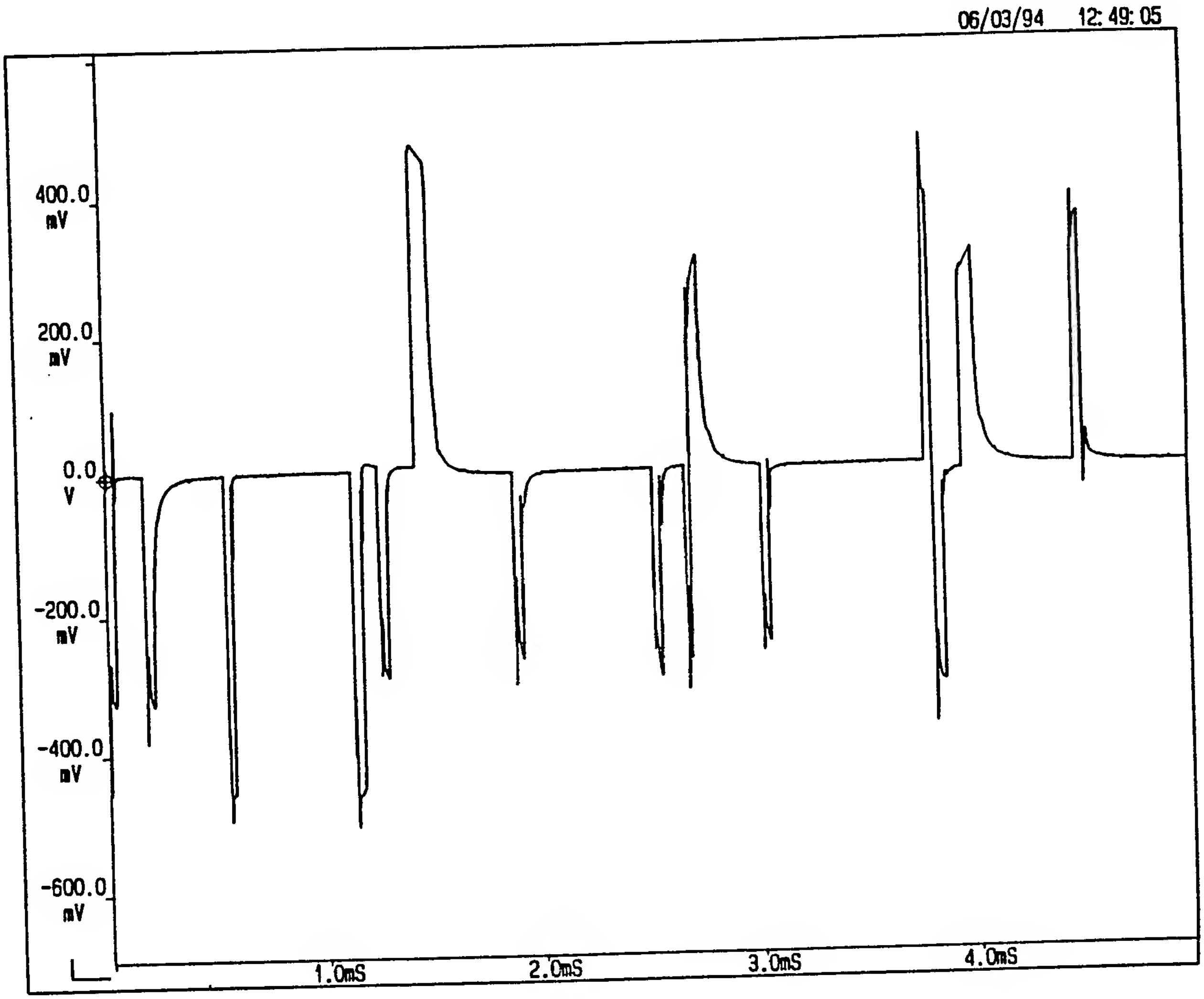


Figure 12: Induced Voltage Waveform for the Yoke at No-Load

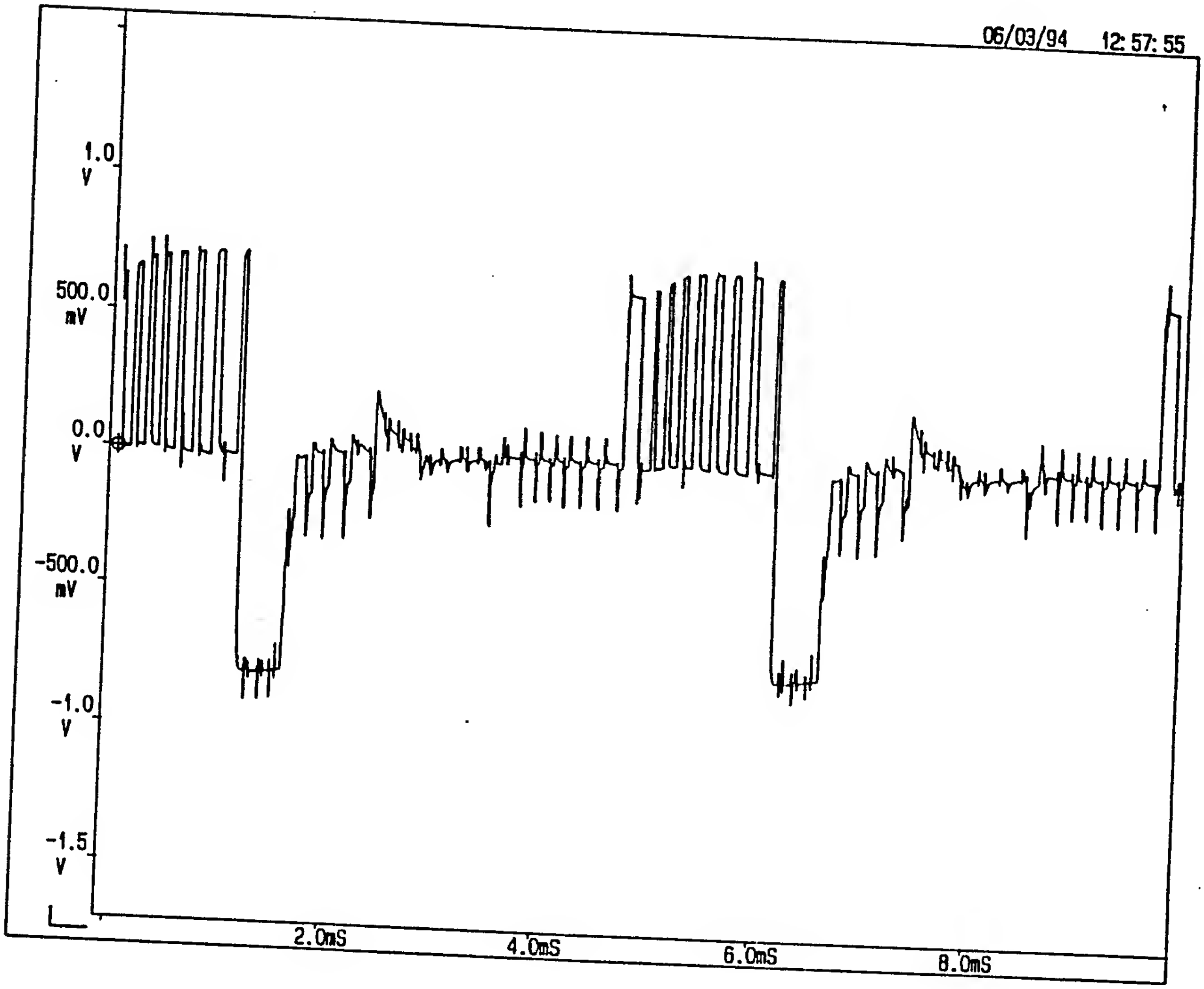


Figure 13: Induced Voltage Waveform for Phase One Coil at 40% load and 2000 RPM

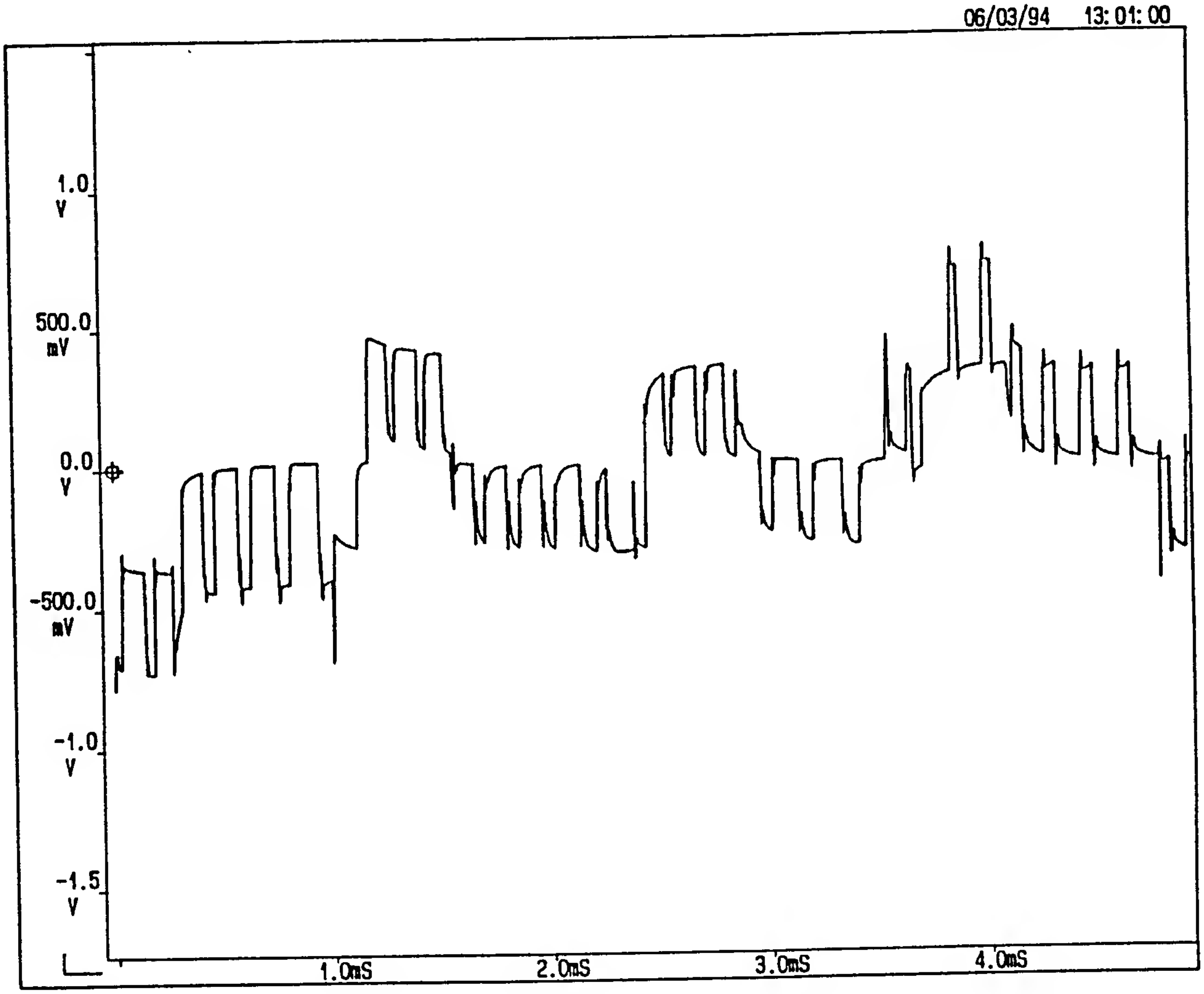


Figure 14: Induced Voltage Waveform for Yoke Coil at 40% Load and 2000 RPM

waveform for the yoke section. Consequently, to theoretically calculate the core losses of the switched reluctance motor is a very challenging task.

**LOW SPEED OPERATION**

The induced voltage waveforms for phase one and yoke search coil for 500 RPM and 75% of the rated load are shown in Figures 15 and 16 respectively. It is still in the current chopping mode. The yoke induced voltage waveform is much better compared to those of earlier cases as far as the distinct pattern is concerned.

**ON GOING RESEARCH AREA AND CONCENTRATION**

(1) Temperature Profile and Thermal Modeling

Three temperature sensors, LM335Z from the National Semiconductor, were installed inside the motor when the motor was built. The locations chosen are the phase winding, pole, and yoke section. Unfortunately, the pinout for these three temperature sensors are not identified at all. Significant amount of time was spent in determining the pinout in order to design a circuit to read the voltage which in turn will yield the temperature at the winding, pole and yoke section of the switched reluctance motor. So far, the pinout for two out of the three temperature sensors are identified. The last one either can not be determined accurately or it is damaged during the manufacturing process. Therefore, it is an on going process to obtain the temperature profile at different locations of the switched reluctance motor.

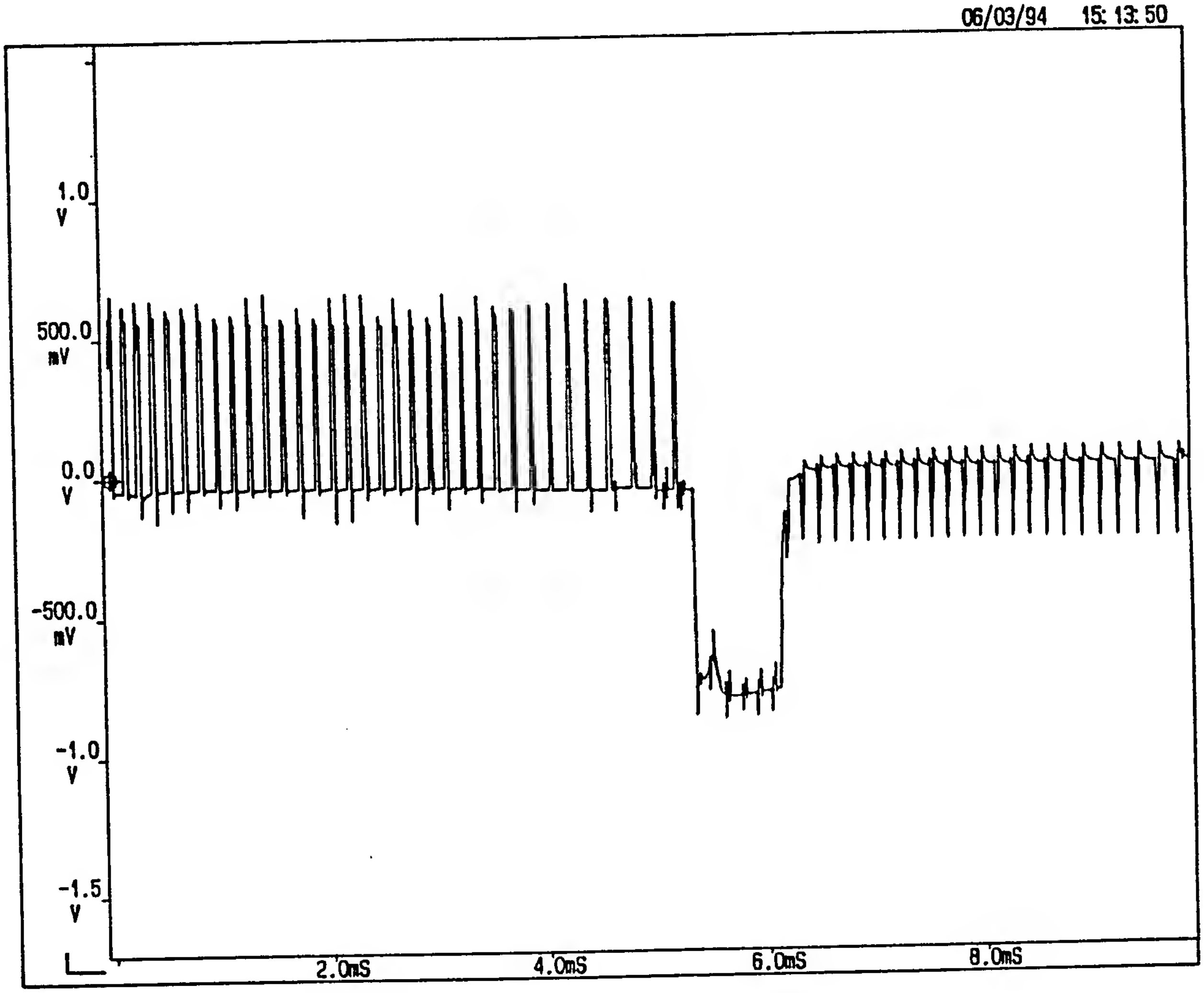


Figure 15: Induced Voltage Waveform at 75% of Rated Load and 500 RPM for Phase One Search Coil

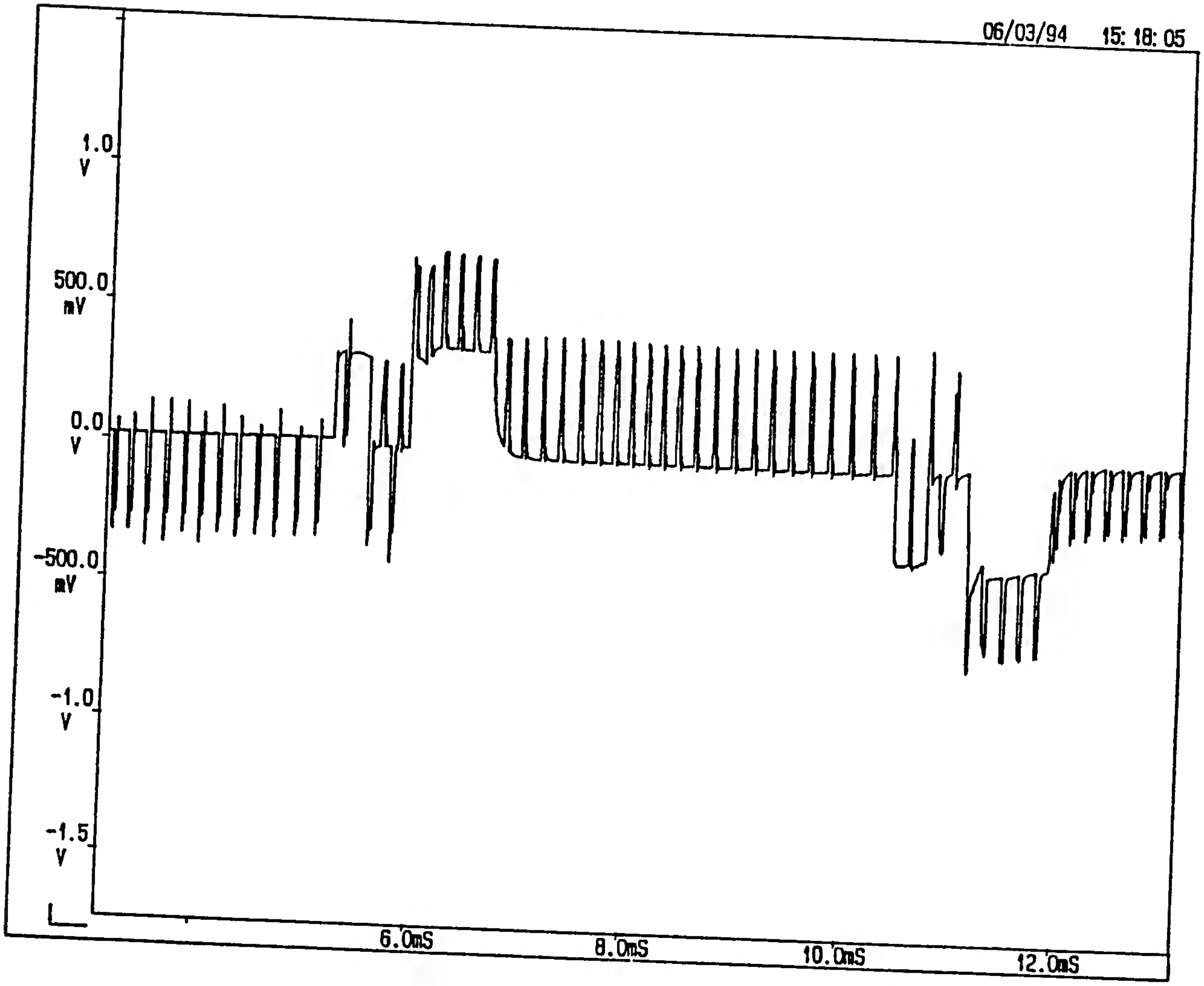


Figure 16: Induced Voltage Waveform for Yoke Search Coil at 500 RPM and 75% of Rated Load

(2) Core Losses Calculation and Prediction

In order to calculate the core losses, accurate flux waveform at different location of the switched reluctance motor is essential. Presently, effort is spent on getting flux waveforms from the induced voltage waveforms. Once this process is completed, core losses calculation can proceed with ease as shown in reference [5].

(3) Thermal Modeling of Switched Reluctance Motor

After accurate core losses data are obtained, effort will be spent to get a decent and simple thermal model using resistance and capacitor elements as shown in reference [6].

**CONCLUSION**

(1) The inductance measurement for the search coil wounded on the yoke section does not have significant changes when the rotor position varies. Thus it is not useful in identifying the rotor position.

(2) The inductance measurement from the search coils wounded on each pole faces do vary cyclically when the rotor poles change position. It will be better to utilize the slope of those inductances instead of absolute values in order to identify the rotor position.

(3) With the limited no. of stator poles, the resolution which can be achieved through the search coil might be very course or limited.

(4) The induced voltage and deduced flux linkage waveforms for the yoke section is quite irregular in nature. This makes the core losses calculation very difficult.

(5) The induced voltage and deduced flux linkage waveforms

for the stator poles are very predictable and hence pose less problem.

## REFERENCES

[1] Lawrenson, P.J., Stephenson, J.M., Blenkinsop, P.T., Corda, J., Fulton, N.N., "Variable Speed Switched Reluctance Motors, *IEE Proceedings*, Vol. 127, Pt. B, No. 4, July, 1980.

[2] *Electrical Machinery*, Fitegerald, A.E., Kingsley, C. Jr., Umans, S.D., McGraw Hill, Fifth Edition, 1990.

[3] Krishnan, R., Bharadwaj, A.S., Materu, P.N., "Computer-Aided Design of Electrical Machines for Variable Speed Applications," *IEEE Transactions on Industrial Electronics*, V. 35, No. 4, Nov. 1988, pp. 560-571.

[5] Lavers, J.D., Biringer, P.P., "Prediction of Core Losses for High Flux Densities and Distorted Flux Waveforms," *IEEE Transactions on Magnetics*, Vol. Mag-12, No. 6, November, 1976, pp. 1053-1055.

[6] Switched Reluctance Motors and Their Control, Miller, T.J.E., *Oxford Science Publications*, 1993.

[7] Vo, Lawrence, "Single-Phase PWM Inverter Using Mos-Controlled Thyristor," Master Thesis, San Francisco State University, 1993.

# Process Migration over a Network of Workstations

Dallas J. Marks and David A. Charley
Graduate Students
Department of Electrical & Computer Engineering

University of Cincinnati
Cincinnati, Oh. 45221-0030

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

March, 1994

# Process Migration over a Network of Workstations

Dallas J. Marks and David A. Charley
Graduate Students
Department of Electrical & Computer Engineering
University of Cincinnati

## Abstract

Although computer performance is increasing at a record pace, there is always a class of users that never seems to be satisfied with current hardware performance. These users often run computationally-intensive programs nearly 24 hours a day and are always looking for additional CPU power. Many organizations have computing resources that go unused during evening and weekend periods. Such machines make ideal candidates to run additional computing jobs.

The Distributed Configuration Manager (DCM) is a graphical utility that allows intensive users to access the additional CPU power of idle workstations on their network. DCM allows users to run processes on idle workstations of colleagues during non-peak usage periods such as evenings and weekends, yet return these machines to their original users during peak usage periods. This is achieved by the use of a process migration mechanism that allows programs to be suspended on one machine and transparently restarted on another machine from its point of suspension. The availability of a process migration mechanism promotes resource sharing and maximum usage of resources, allowing users to access CPU power otherwise unavailable to them.

This report presents the design and development of the Distributed Configuration Manager and its process migration mechanism. The migration mechanism is based on the mechanism found in the Condor Distributed Batch System developed at the University of Wisconsin. Specifically, this mechanism has been modularized and enhanced to support an on-demand migration policy and the migration of processes that use UNIX System V shared libraries. Extending the mechanism to support shared libraries is significant because programs built with shared libraries require less physical resources such as main memory and secondary storage, providing greater overall system performance.

The DCM system has been developed under AFOSR contract number F49620-90-C-09076 at the Crew Systems Integration Laboratory at Wright-Patterson AFB. DCM runs on a network of Silicon Graphics 4D workstations running version 4.0.5 of IRIX, the Silicon Graphics UNIX implementation. The DCM mechanism requires no modifications to the hardware or operating system software. DCM has been designed using object-oriented techniques and coded in the C++ language. The object-oriented design provides an application framework for future enhancements to the DCM environment. Such enhancements might include fault tolerance or load balancing.

## Process Migration over a Network of Workstations

Dallas J. Marks and David A. Charley

# 1 Introduction

Despite the fact that hardware performance continues to rapidly advance, there always seems to be a class of users that are always unhappy with the state of current performance levels. Such users can be labeled as intensive users. Intensive users tend to execute long-running processes, such as simulations, around the clock and they are eager to find additional computing power. In a typical working environment, much of a facility's computing resources are idle during evening and weekend hours. The intensive user may choose to run programs on other users' processors during these idle periods; however, the user is usually forced to return the machines to their original state when their normal users return. This poses a significant problem if the computation-intensive processes do not complete in a reasonable amount of time. Normally, a long-running process would have to be terminated without yielding any results.

The Distributed Configuration Manager (DCM) is a configuration utility that allows the distributed processing resources of a network to be managed by both casual and intensive users of that network. Its main feature is a process migration system that allows computing jobs to be moved from machine to machine to allow users to take advantage of idle computing resources. A user-friendly graphical user interface (GUI) is used to control network configuration and process migration.

Process migration refers to a transparent mechanism that allows programs to be halted in the middle of execution on one system and transparently moved to another system and restarted there. Such a mechanism allows computation-intensive jobs to be started on idle systems, transparently moved when these systems become loaded by their intended users, and moved back again when the machines become idle again. The ultimate achievement is that the computer resources of an organization can be fully utilized, providing intensive users with additional power that cannot normally be found in their hardware budgets.

## 1.1 Motivation for Research

Process migration is not a new idea. It has been implemented in several research systems to support advanced features such as resource sharing, fault tolerance and load balancing. Unfortunately, most process migration mechanisms are built into custom operating systems such as the V System [21], Charlotte [2], and Sprite [7]. Using a custom operating system restricts the number of machines that are available to run jobs; most users in an organization rely on a standard operating system such as UNIX to run commercial productivity applications and are not motivated to use a new operating system. However, if process migration can be achieved on top of an existing operating system such as UNIX, average users may continue to run their applications and intensive users can use process migration to "borrow" their workstations.

Several process migration systems have been implemented in UNIX [20]; however, most have required modifications to the kernel and therefore are in the same class as custom operating systems; ordinary users will not choose to use such an operating system. The Condor Distributed Batch System [11, 9, 10, 3, 12, 14] is unique because it supports process migration completely outside of the UNIX kernel, i.e. on top of an unmodified operating system. The Condor system is very powerful; however, it has certain restrictions that prevent it from being useful in application areas such as distributed simulation. These restrictions include the inability to migrate communicating processes and processes that use shared libraries.

The goal of the research presented here has been to extend the types of processes supported by Condor. Although our intentions are to develop a generic process migration system that works with any process or set of processes, our long-term research goals are focused toward constructing a process migration system that supports the QUEST Distributed VHDL Simulator.

## 1.2 The QUEST Distributed VHDL Simulator

Because of the sheer size and complexity of planned designs, the Cockpit Avionics Office at Wright-Patterson Air Force Base needs a powerful VHDL simulator to aid in the design of specialized hardware for cockpit display generators. These cockpit display generators will be used in next-generation transports and fighters and in a proposed retrofit to the Advanced Tactical Fighter [16]. The QUEST VHDL Simulator is a high-performance distributed simulator originally written for execution on an Es-Kit multiprocessor at the University of Cincinnati under DARPA-supported research [6]. Distribution and the resulting parallel execution of multiple objects provides greater performance than can be achieved by running simulations on a single processor. The QUEST simulator has recently been ported to a network of Silicon Graphics 4D workstations [15]. The main features of this port include shared libraries for code size reduction, and a unique message-passing interface that transparently exchanges messages between objects using a hierarchical delivery system of local shared memory, ethernet and SCRAMNet, a high-speed shared memory networking system [5].

A typical QUEST simulation may contain tens or even hundreds of simulation objects (UNIX processes) distributed on a network. Due to shrinking defense budgets, the Cockpit Avionics Office VHDL design lab has only a handful of UNIX workstations that are capable of running QUEST. However, many other workstations are connected via a network. These additional machines are used for real-time flight simulation studies but are often unused during evening and weekend hours. A process migration system that supports QUEST will provide maximum use of network resources during these non-peak usage periods, yet also guarantee that simulations run until completion when fewer machines are available. The ultimate goal of the Distributed Configuration Manager is the capability to migrate QUEST simulator objects. The current version of the Distributed Configuration Manager supports migration of processes using shared libraries. The current version also supports migration of QUEST processes that interface to the hierarchical communication system.

## 1.3 Terminology

Several terms found throughout this report are defined here.

Process migration is a mechanism that moves a process from one computer to another during execution without loss of state. The machine upon which a process executes is known as the source or host machine. A process is migrated to a new machine known as the target.

Migration is performed by the Distributed Configuration Manager in a three-step sequence of process suspension, checkpointing and restarting. Process suspension refers to the act of commanding a process to halt its execution. Checkpointing is a procedure that saves the suspended process state for later execution. Our system preserves process state in a file, known as the checkpoint file. Restarting refers to the act of starting the checkpoint file on a new host machine. Upon restart, a process begins execution from the point of its original suspension.

On-demand process migration refers to manually-controlled process migration that occurs only at the user's request or according to rules defined by the user. These rules constitute what is known as a policy, specifying which processes to migrate and where to restart them. The Distributed Configuration Manager uses an on-demand mechanism that is user-initiated with a graphical user interface (GUI).

While the Distributed Configuration Manager uses a user-driven policy, it is possible to construct automated policies for controlling the process migration mechanism. Two examples of automated policies are fault tolerance and load balancing. Fault tolerance is a characteristic of a system indicating that it is immune to failure. A fault tolerance policy would be responsible for identifying machines that are about to shut down due to a fault, checkpointing all processes, and either migrating these processes to new hosts or restarting them on the same machine after it is restored. Load balancing refers to dividing processing requests equally over a set of machines, usually machines connected to each other on a network. A load balancing policy would be responsible for determining which machines in a set were either too heavily loaded with processes or too lightly loaded. The policy would use process migration to redistribute processes on heavily loaded machines to lightly loaded machines to insure that each machine in the set was equally loaded.

## 1.4 Report Objectives

This report describes the design, implementation, and evaluation of Distributed Configuration Manager, an on-demand process migration facility for UNIX networks. The main objective of this report has been to demonstrate that process migration of processes using UNIX shared libraries is possible on one system and to develop concepts to aid in mapping our experiences to other systems. A secondary objective of this report has been to use object-oriented design and software construction techniques to provide an application framework that will allow the Distributed Configuration Manager to be easily extended.

## 1.5 Report Organization

The following section, section 2, discusses previous work in the area of process migration. Of particular interest is the Condor Distributed Batch System [3], whose checkpointing mechanism forms the basis of the Distributed Configuration Manager mechanism. Section 2 also discusses background information on the underlying UNIX system primitives that allows process migration to be performed.

Section 3 provides a detailed description of the migration system, which includes the Policy Manager and Migration Server. In particular, Section 3 illustrates how the Condor migration mechanism has been extended to support shared libraries.

Section 4 provides an analysis of the Distributed Configuration Manager and the achieved objectives. Finally, Section 5 outlines several directions for continuing research.

# 2 Background and Related Work

Process migration is defined as "the transfer of a sufficient amount of a process's state from one machine to another for the process to execute on the target machine." [20] It has been used in many distributed systems, usually to implement fault tolerance and load balancing.

This section will outline the design considerations that must be addressed in designing a process migration mechanism, provide a brief history of existing systems, give an overview of the Condor Distributed Batch System, and analyze how the UNIX operating system stores process state.

## 2.1 Design Considerations

The designers of the Sprite operating system established four design considerations for their system [7]. These considerations should be satisfied for any process migration system and our goal has been to satisfy these considerations for our process migration system in the Distributed Configuration Manager.

- Transparency

  A process migration system should have a network-transparent execution environment. Such an environment will allow a process to execute on any machine in the network without restrictions. Achieving transparency may result in design tradeoffs. For instance, a process that uses a specific hardware device, such as a graphics terminal, may be unable to migrate because the hardware resource is only available from a particular machine.

- Minimal Interference with the System

  Process migration should not introduce excessive interference with either the process being migrated or the system as a whole. The migration mechanism should operate atomically.

- Residual Dependencies

  The process migration mechanism should be designed to minimize residual dependencies on previous locations. For instance, once a process has been migrated from Machine A to Machine B, the process should no longer require Machine A.

- Complexity

  Complexity is an important factor in a custom operating system such as Sprite. Process migration tends to affect virtually every major piece of an operating system kernel [7]. The migration mechanism tends to be complex even when migration is implemented outside of the kernel.

Not all process migration systems fully adhere to these four design considerations. In practice, it is not often possible to fully satisfy all of them. A series of tradeoffs must be made according to the goals of a particular process migration system. For instance, our implementation requires that all migration mechanisms exist outside of the operating system kernel. The Distributed Configuration Manager's migration mechanism is transparent and free of residual dependencies, but at the expense of increased complexity.

## 2.2 Other Implementations of Process Migration

Several systems have been designed for process migration. Each of these systems has its own design constraints; however, process migration mechanisms fall into two broad categories. These are:

- process migration inside the operating system kernel
- process migration outside the operating system kernel

This is an important distinction because one of our requirements dictates that the Distributed Configuration Manager performs migrations on workstations with unmodified kernels. Our design alternatives are automatically restricted by this requirement.

This section presents the Charlotte [2] and Sprite [7] operating systems as examples of process migration implementations inside the operating system kernel. The Condor Distributed Batch System [3] is presented as an example of a process migration implementation outside the operating system kernel. Both Smith [20] and Ankola [1] offer more complete surveys of process migration systems; the purpose here is to show the differences in design approach when kernel data structures are available to the migration mechanism (Charlotte and Sprite) and when they are not (Condor).

### 2.2.1 Sprite

Sprite is an operating system for a collection of personal workstations and file servers on a local area network [7]. Typical Sprite applications include parallel compilation and simulation. The motivation for adding process migration capability to Sprite was the existence of idle machines on the network that could be freely used for additional computational power.

Sprite's overriding design objective was to provide transparency to the user. Transparency in Sprite means that a process's behavior is not affected by migration. Its execution environment appears identical on any machine on the Sprite network; processes have global access to resources such as files and devices. An additional benefit of transparency is that a process's appearance to the rest of the world is not affected by migration. Unlike process identifiers in systems such as UNIX, Sprite processes use global identifiers. The system and its users always see processes running on their original hosts, even if such processes are executing remotely. For instance, a user need not log in to a remote machine to halt a process. Because the process appears to be executing on the user's desktop workstation, the process may be killed directly on the host console.

A database of idle processors is maintained by a central migration server. Load-average daemon processes on each machine notify a central migration server when machines are determined to be idle. When a user requests that a process run remotely, this central migration server selects a host machine and executes the process there.

Although the selection of idle hosts is automatic, the migration policy is determined by the user. Processes are initiated from a home machine, typically on a user's desktop. A home machine is the machine where a process would execute if there were no migration at all. The user may request migration during process startup (remote invocation) or during execution. The Sprite central migration server automatically assigns a remote host. Should this host become busy, the process is evicted from the host and migrated back to its home machine, where it continues to execute until a new idle host is found.

The Sprite migration mechanism operates by terminating a process and storing dirty pages of the process virtual address space in a special file. When the process is restarted on a new host, it will retrieve these dirty pages as the process executes, using a form of lazy-copying. Many migration systems transfer the entire virtual address space to the new processor before the process continues execution. Because Sprite delays the loading of the virtual address space, it can achieve better performance.

The Sprite operating system has minimized residual dependencies but not eliminated them. Although processes leave no residual dependencies on remote hosts, all processes have a residual dependency on their home machine. Some kernel calls achieve transparency by forwarding requests to the home machine, such as gettimeofday(). Because of the residual dependency on the home machine, users are unable to migrate processes to a new home machine in case of failures. However, the Sprite system designers felt that achieving transparency was more important than providing reliability.

### 2.2.2 Charlotte

Charlotte is a message-based distributed operating system designed for experimentation with distributed algorithms and load distribution strategies [2]. Process migration facilities have been added to the operating system to better support such experiments.

Because it is a research system, Charlotte uses no specific policy for migration. Instead, it has been designed with the migration policy separated from the underlying migration mechanism. The migration mechanism is located in the kernel; it supports concurrent multiple migrations and premature cancellation of migration. Policies are run as utilities and can be easily created and modified to support research ideas.

Unlike Sprite, that uses a variation of lazy copying to transfer its virtual address space, Charlotte freezes the process and sends the entire virtual address space to the target machine. This method is easy to implement; however, it takes several seconds to perform.

One of Charlotte's features is fault tolerance and the migration mechanism has been designed to maintain this feature. The migration mechanism leaves no residual dependencies on the host. Therefore failure of the source host does not affect the process unless it is communicating with a process on that host.

### 2.2.3 The Condor Distributed Batch System

The Condor Distributed Batch System is a process migration facility designed and developed at the University of Wisconsin. It first became operational in 1984 when it ran on a network of VAX 11/750 minicomputers. Over the past few years Condor has been ported to ten different hardware platforms including Sun3, Sun Sparc, HP PA-RISC, IBM RS/6000, DEC DecStation, and Silicon Graphics 4D UNIX systems. As of this writing, a port of Condor to DEC Alpha machines and a version of Condor that runs on top of the PVM parallel programming environment, Condor-PVM, are under development [13].

The designers of the Condor system identified three basic categories of computer users: casual users, occasionally heavy users, and heavy users. Heavy, or intensive, users are "people who frequently do large numbers of simulations, or combinitoric searches. These people are almost never happy with just a workstation, because it really isn't powerful enough to meet their needs" [3]. Unlike the casual and occasionally heavy users, intensive users often keep their machines busy 24 hours a day.

Condor is an attempt to use the available power from casual and occasionally heavy users to satisfy the needs of heavy users. Condor software monitors the activity on all participating workstations in the local network. Machines that are determined to be idle are placed into a resource pool, or processor bank. The bank is a dynamic entity; workstations enter the bank when they become idle and leave when busy.

#### Features of Condor

Condor has been designed with the following seven design characteristics [3]:

1. Condor does its work completely outside the UNIX kernel. This allows users of Condor to coexist with other users who rely on UNIX.

2. No special programming is required to use Condor. Condor is able to run ordinary UNIX programs, only requiring the user to re-link, not recompile them or change any code.

3. The local execution environment is preserved for remotely executing processes using a shadow process mechanism. A shadow process runs on the local machine and executes operating system calls on behalf of a remotely executing process. Shadow processes are necessary if the remote environment is not identical to the host environment. For example, machines in different time zones or with different file systems require the shadow process mechanism.

4. The Condor central manager is responsible for locating and allocating idle workstations. Condor users do not have to search for idle machines, nor are they restricted to using machines only during a static portion of the day.

5. Owners of workstations have complete priority over their own machines. Condor remotely executes processes only when an idle machine is found. When an workstation's owner returns to reclaim his or her workstation, Condor will automatically move all remote processes back to the host machine or to another idle machine. Condor's operation is transparent to other system users.

6. Users of Condor may be assured that their jobs will eventually complete. If a user submits a job to Condor which runs remotely but is not finished when the workstation owner returns, the job will be checkpointed and restarted as soon as possible on another machine.

7. File systems of remote execution sites are untouched by remotely executing jobs, preventing Condor from cluttering up private disk space. This problem can be eliminated if a transparently distributed file system, such as NFS, is used.

### Limitations of Condor

Although Condor is a powerful environment, it does possess limitations [3, 14]. These include:

1. Migration is limited to single-process jobs; programs that use the fork(2), exec(2), and similar calls cannot be migrated.

2. Signals and signal handlers are not supported; programs that use the signal(3), sigvec(2), and kill(2) calls cannot be migrated.

3. Processes using interprocess communication (IPC) cannot be migrated; the socket(2), send(2), recv(2), and similar calls cannot be used.

4. All file operations must be idempotent. Read-only and write-only file accesses work correctly, but programs which both read and write the same file may not work. In addition, memory-mapped files are not supported.

5. Each Condor job has an associated checkpoint file that is approximately the size of the process address space. Enough space to store the checkpoint file must be available both on the host and remote machines.

6. Processes that use shared libraries cannot be migrated.

Many programs do not require these advanced operating system features; however, these features should be supported by a process migration mechanism if it is to be considered universal. The Distributed Configuration Manager demonstrates that the Condor migration mechanism can be extended to support shared libraries. It is conceivable that extensions could be made for better file and multi-process support.

### 2.2.4 Discussion

Because of the overriding requirement that the operating system kernel may not be modified, the Condor Distributed Batch System has been selected as the basis of the Distributed Configuration Manager migration mechanism. Migration outside the kernel presents some challenges, particularly because kernel data structures that store process state cannot be accessed or modified. Lack of kernel modification also limits the amount of system tuning that can increase performance. However, it is apparent from the analysis of Sprite and Charlotte that even migration systems in the kernel do not solve all performance problems.

## 2.3 Basic Components of Process State

The purpose of this section is to identify the generic components of a process state and how they are stored. The process state must be identified before a process migration facility can be constructed. In general, process migration involves three steps [1]:

1. Suspension of a process on its original (source) machine.
2. Transfer of process state to a new (target) machine.
3. Resumption of execution on the new host.

The crucial step in any process migration mechanism is recovering and transferring the state of the suspended process. The process migration mechanism used by Condor and modified for the Distributed Configuration Manager is unique because it operates on top of the UNIX kernel. Because the migration mechanism is not built into the operating system, process state must be obtained without illegal access to internal kernel data structures. When the process state may not be determined directly, the migration designer can either leave the state on the source machine and forward the state to the target, or use a comparable state on the target, sacrificing transparency. The designers of the Sprite operating system [7] identify five components of the state of an abstract computer process:

- Virtual memory
- Process execution state
- Open files
- Interprocess communication
- Other kernel states

These five states are examined in detail below.

### 2.3.1 Virtual Memory

Virtual memory refers to both the program instructions (text) and data that a program uses during execution. The text and data make up the bulk of process state. In UNIX systems, the program text is stored in an a.out or executable file. Most UNIX implementations use a variation of the System V COFF file format. In addition to text, the executable file also contains initialized data prior to system execution; however, the state of program data at the point of program suspension is stored by the UNIX system in a core file. Virtual memory under UNIX will be discussed further in Section 2.5.

### 2.3.2 Process Execution State

The process execution state refers to the information that identifies the point in the virtual address space where a process is currently executing. The storage of this data is generally dependent on the underlying physical hardware upon which a process is running. Process execution state data generally includes the process control blocks, the hardware registers, the stack pointer, the program counter, and the condition codes. This information is needed when a process is saved and restored during a context switch. Although this information is hardware specific and the Distributed Configuration Manager's migration mechanism permits no illegal access to kernel data, UNIX provides a set of user-level routines for preserving context: setjmp() and longjmp(). The setjmp() system call preserves the process execution state and longjmp() system call restores a state preserved by a previous call to setjmp(). The benefit to the process migration designer is that although these system calls deal with system-specific data, they must be ported by the OS vendor as part of the UNIX standard. The use of these functions may be found in any UNIX systems programming guide and are not discussed here.

### 2.3.3 Open Files

According to Ankola, "information about open files is one of the most difficult [his emphasis] pieces of information to transfer" [1]. A process must store state information for each open file. The state of an open file includes the file identifier, file access pointers, and any cached blocks of data. File state information can be stored in the process virtual address space or hidden in the kernel. To provide uninterrupted access to files, either the files must be copied to the target with current state information, the file operations must be forwarded to the source, or both sites must share a common file system such as NFS. Condor uses both shadow system calls to forward file requests back to the source and maintains a table of open file descriptors that is restored as part of the migration sequence. Ankola chose not to support open files in process migration with apE [1] and the current implementation of the Distributed Configuration Manager also ignores open file support. Supporting open files increases the complexity of the process migration and reduces overall performance of the migration system. However, the open file mechanism could be added to the basic system at a later date if such support is considered mandatory.

### 2.3.4 Interprocess Communication

The state of communicating processes is difficult to define. At any given time, a process may be sending communication data, receiving communication data, or not communicating at all. Migration of such processes is difficult because most of the communication state is known only to the operating system kernel. In addition, a process that is migrated while sending or receiving messages may not be recovered properly because the entire message cannot be retrieved. Because the state of the communication system is difficult to obtain from outside the kernel, migration systems such as Condor do not support communication. However, many research operating systems have been constructed with process migration as a goal; however, designers of such systems have the luxury of designing and maintaining kernel data structures to aid in this task.

Migration of communicating processes outside the kernel can be achieved through the use of a user-level communication package. Unlike operating system communication mechanisms, such as sockets and shared memory, a user-level communication package provides the migration system designer with access to its own internal states. In addition, the user-level package can be modified, unlike the operating system. Migration of communicating processes outside of the kernel has been demonstrated in an enhanced version of the apE graphics toolkit developed at the University of Cincinnati [1]. The apE (animation production Environment) is a software toolkit developed by the Ohio Supercomputer Graphics Project in collaboration with the Ohio State University.

### 2.3.5 Other Kernel States

The kernel typically stores data associated with each process such as the process identifier, user identifier, current working directory, environmental variables, signal masks and handlers, and resource usage information. In the design of a process migration system that functions outside of the kernel, this state data is unavailable. Systems such as Condor ignore this kernel state data and create a new corresponding state on the target machine. This results in some loss of transparency. For example, a migrated UNIX process will be assigned a new process identifier by its new host that is different from the value assigned by the previous host. Transparency is lost because any other process that needs to access the migrated process must be made aware of the new process identifier. In contrast to UNIX, transparency for process identifiers is maintained in the Sprite operating system because process identifiers are maintained globally for all machines; a process maintains a consistent process identifier during its execution.

## 2.4 Process State in the UNIX Operating System

UNIX possesses several standard mechanisms for maintaining process state both on disk and in main memory. Because open files and interprocess communication are not supported in the Distributed Configuration Manager migration system, attention is focused mainly on the process virtual address space. This section discusses the UNIX System V COFF format for storing programs, the core format for storing program debugging data, and how both formats support shared libraries. The COFF and core formats, together with the program state that they preserve, form the backbone of the checkpointing mechanism discussed in section 4.

### 2.4.1 The COFF File Format

COFF stands for Common Object File Format and is the formal definition for the structure of all UNIX System V machine code files [8]. The COFF definition describes a complex data structure that represents object files, executable files, and archive files. Silicon Graphics calls their UNIX implementation IRIX. The object file format used by the IRIX operating system is an enhancement of the basic COFF file format of UNIX System V.

### The Basic Elements of COFF

A computer program typically contains three kinds of information: executable machine code called text, initialized program data, and uninitialized program data.

The text is stored separately from the data to allow an operating system to allow multiple executions of the program to share it. Its isolation from the program data also allows the operating system to protect the text from modification or corruption (write-protection). Initialized data contains values that are set to specific values prior to execution of a program. The initialized data is not write protected like the text; it is available for both reading and writing during the execution of the program. Uninitialized data can also be modified, but by definition such data is not initialized to specific values. Uninitialized data is not stored in a program file to save disk space; however, the operating system must know how much space to reserve for uninitialized data.

The COFF format organizes all three information types into individual areas called sections. They are named as follows:

- The text section contains machine instructions
- The data section contains initialized data
- The bss section contains uninitialized data

The text section contain executable machine code and the operating system treats it as write protected. The data section contains initialized program data and is readable and writeable. The bss section does not actually store data (because this data has no initial values) but instead stores the size of uninitialized data. It tells the operating system how much virtual address space to reserve for the executing process. The bss section is generally made contiguous with the data section when the program is loaded into virtual memory. All UNIX systems initialize the bss section to zeros when it is created in virtual memory. The bss acronym comes from IBM mainframe terminology; bss means memory Block Started by Symbol, a block of memory that is not initialized.

The COFF definition also specifies a symbol table and a string table. The symbol table is used by high-level languages, such as C or C++, to store symbols such as procedure and variable names. These symbols are invaluable during the debugging process. The string table is used in conjunction with the symbol table. It defines symbols that exceed the eight-character limit of the symbol table format. Because the symbol and string tables are only useful during debugging, most UNIX implementations provide utilities to strip the tables from an executable file to reduce the program file size. This operation is generally performed after program development.

### The Benefits of COFF

The COFF definition provides the UNIX system with two major benefits: enhanced portability, and system extensibility.

The COFF definition enhances portability because it provides an abstraction between the specifics of a hardware platform and the basic fundamentals of a software program. Machine instructions vary from machine to machine and even data can be stored in different formats (such as little-endian and big-endian). COFF minimizes and localizes the amount of machine dependent code in different ports of the UNIX operating system. Most of the porting work of the UNIX operating system involves changes to the C compiler code generator, assembler, debuggers, and a few localized areas in the kernel such as the program entry/exit, system call service, and interrupt tables [8]. Enhanced portability also aids process migration, as will be explained shortly. Because all UNIX variants follow the basic COFF format to varying degrees, much of the code needed to perform migration can be reused when porting the process migration mechanism to different UNIX platforms.

The COFF definition also provides a framework that allows for system extensibility. Additional sections may be added to the basic COFF definition to implement features exclusive to a particular version of UNIX or to take advantage of the underlying hardware architecture. For example, the Silicon Graphics IRIX object format [17] stores initialized data in three data sections rather than one. These are the read only data section (rdata), large data section (data) and small data section (sdata). The IRIX object format also stores the bss data in two sections. These are the block started by storage section (bss) and small block started by storage section (sbss). This additional subdivision of sections allows the linker to localize data by its type to enhance system performance. A diagram of the IRIX extended COFF format is shown in Figure 2.2.

The COFF definition allows for generic extensions to its basic format and has been modified by various UNIX vendors to add shared library capability or machine-independent data formats. Unfortunately, while enhanced portability aids in the portability of the checkpointing mechanism of process migration, the ability to enhance the COFF definition limits the portability of the checkpointing mechanism and requires specific changes for each platform supported.

### 2.4.2 The core File Format

The UNIX system writes out a core image of a terminated process when any of various errors occur. The most common errors are memory violations, illegal instructions, bus errors, and user-generated quit signals. The process of generating a core image is known as a core dump, a holdover term from early mainframes that used magnetic core memory. The purpose of the core image is to store the state of a process when it terminated into a file. A program called a symbolic debugger can then examine both the original program and the core image and allow the user to determine what went wrong.

On IRIX systems, the format of the core image is defined by <core.out.h>. The core format for IRIX was designed by Silicon Graphics and does not adhere to a pre-existing standard format. It consists of a header, maps, descriptors, and section data.

The header data includes the process name, the signal that caused the core dump, the descriptor array, and the corefile location of the map array.

Each descriptor defines the length of useful process data. For example, one descriptor defines the general-purpose registers at the time of the core dump. Each map defines the virtual address and length of a section of the process at the time of the core dump. The map data is present in the core image at the file location given in the descriptor only if the VDUMPED flag is set in the map. The process stack and data sections are normally written in the core image while data available in the COFF file such as standard text and shared library text is not. IRIX specifies eight different possible map types that may be stored in the core:

- VTEXT text map (not normally present in core)
- VDATA data/bss map
- VSTACK stack map
- VSHMEM shared memory map
- VLIBTXT shared library text map (not normally present in core)
- VLIBDATA shared library data map
- VGRAPHICS graphics hardware map
- VMAPFILE memory mapped file map

Most of the data that must be preserved is found in the core image. For instance, the data/bss map contains part of the virtual address space of the process. The stack map contains the process execution state. The shared library data map will be used by our enhanced checkpointing mechanism to restore the state of any shared libraries used in a process. Although Condor and the Distributed Configuration Manager do not support shared memory or memory memory mapped files, the state data is available in the core image. Mechanisms may be devised to restore this data, in the same way that existing mechanisms restore data/bss, stack, and shared library data.

### 2.4.3 UNIX System V Shared Libraries

Because the UNIX system is being used on smaller computer systems, it is increasingly important to use disk space, memory, and computer power more efficiently. A shared library can offer savings in all three areas.

In the UNIX system, common code libraries are typically stored in archives. An archive is a collection of object files that are combined into a single file. When a user wishes to use library functions, he or she merely tells the linker to include this single library rather than a large and cumbersome collection of object files. A good example is the C language library archive, libc.a. Nearly every executable file contains some, if not significant, portions of the C library. Additional examples of archived libraries on Silicon Graphics systems include the GL graphics library and the Font Manager library. Again, these are libraries that are used by a significant number of programs.

Improved system performance can be achieved by storing common libraries such as the C library in a manner such that only one copy exists on disk and in physical memory using a shared library. A shared library is "a file containing object code that several a.out [executable] files may use simultaneously while executing." [18] The current shared library implementation in IRIX is based on System V Release 3 specifications.

#### Advantages of Shared Libraries

A shared library offers several advantages over simple archives by not copying code into individual executable files. It can:

- save disk storage space

  Because shared library code is not copied into all the executable files that use the code, programs built with shared libraries are smaller and use less disk space. This not only saves space but requires less I/O activity, a major performance penalty in any computer system.

- save memory

  By sharing library code at run time, the dynamic memory needs of processes are reduced. Again, I/O activity is reduced by reducing both paging and swapping activity.

- make executable files using library code easier to maintain

  Because shared library code is loaded into a process' address space at run time, a shared library may be updated without requiring updates to all of the processes that use the shared library. Such updates are not possible with standard archives; updating a standard archive will require all programs using the archive to be relinked.

#### Organization and Operation of Shared Libraries

A shared library consists of two parts: the host library and the target library. The host library is just like a standard archive library. Each of its members, typically a complete object file, defines some text and data symbols in its symbol table. The link editor searches this file when a shared library is used during the compilation or link editing of a program and includes relevant portions into the executable. The target library closely resembles an executable file. This file is read by the operating system if an executing process needs a shared library.

An executable linked with a shared library will contain a special section called lib that defines which shared libraries are needed. When the program is executed, the operating system will use this section to load the appropriate library code into the virtual address space of the process. This makes all required library code available to the process prior to execution.

Shared libraries enable the sharing of text sections in the target library, which is where text symbols are defined. Although processes that use the shared library have their own virtual address spaces, they share a single physical code for each process that attaches a shared library's text. The target library cannot share its data sections. Each process using data from the library has its own private data region (contiguous area of virtual address space that mirrors the data section of the target library). Processes that share text do not share data and stack area so that they do not interfere with one another. The target library is very similar to a standard executable file; it can share its text but not its data. Like any other UNIX file, the shared library target has access permissions; a process must have execute permission to use a target library.

Symbols in the shared library are statically linked to the target executable during the link phase; however, a shared library contains a branch table to provide a level of indirection between the symbols in the host library and the target executable. A branch table associates text symbols with absolute addresses that do not change, even when library code is changed. Each address labels a jump instruction to the address of the code that defines a symbol. The use of the branch table allows the shared library to be updated for bug fixes or the addition of additional functions without recompiling applications that use the shared library.

When an executable is built using a shared library, it will contain two additional sections to the standard COFF definition, the lib and init sections. The lib section contains only relocation information: the addresses of the referenced routines. Unlike the standard COFF definition, the init section is always present in the Silicon Graphics extended COFF definition. The init section is the only part of the shared library code that becomes part of the executable file. It contains a series of initialization statements. In an executable that uses no shared libraries, the init section will be 32 bytes long. In an executable that uses shared libraries, the init section will be longer depending on the initialization code present in the shared libraries used.

## Building Shared Libraries

A shared library archive is built using the UNIX system tool mkshlib. The IRIX Programming Guide Volume II [18] outlines six major tasks required in the building of a shared library:

- Choose region addresses
- Choose the pathname for the shared library target file
- Select the library contents
- Rewrite existing library code to be included in the shared library
- Write the library specification file
- Use the mkshlib tool to build the host and target libraries.

These tasks are detailed in the system documentation. Several requirements are placed upon the shared library designer.

1. Fixed virtual addresses must be specified that do not conflict with existing shared libraries.
2. The path location of the shared library must be specified (a shared library cannot be moved inside the file system once created).

3. The most difficult step, from practical experience, is writing the library specification file. The process of building a shared library is very tedious and, not surprisingly, is the source of most of the limitations of using shared libraries.

### Limitations of Shared Libraries

Although shared libraries provide many advantages, the current implementation has a number of disadvantages. The first limitation is the requirement that shared library text and data regions be statically bound to a specific virtual address. These addresses are statically bound so that all references to shared library procedures and data may be resolved at link time. However, conflicts can arise if the software developer chooses to use two or more shared libraries in the same application that require the same virtual address space. In this case, one or more of the shared libraries must be rebuilt to exist at an unused address range.

This problem is alleviated somewhat by the second limitation of shared libraries; they are extremely difficult to build. The mkshlib tool requires that each external symbol (such as a C procedure name or piece of data) be defined in a branch table that provides a layer of indirection between a library reference in an application and the actual location of the reference in the shared library. The branch table is provided as a convenience, as a shared library can be modified for enhancements or bug fixes without requiring programs that reference the shared library to be recompiled. However, the creation of the branch table is tedious given the current state of shared library tools. The tedium of shared libraries can be eliminated with better tools.

The third limitation is that the location of the shared library is statically bound to processes that use it. This requires any user shared libraries to be stored in a fixed location. A much more flexible system would allow shared libraries to be physically relocated.

The fourth limitation is the difficulty of using imported symbols. If a function in a shared library uses the printf() function in the standard C library, the printf() symbol must be redefined with a dummy symbol to avoid symbol resolution until the final link step.

Finally, the most significant limitation is the inability to create shared libraries using the C++ language. C++ mangles symbol names during compilation, making them difficult to recognize during symbol resolution. This difficulty could be averted if a shared library tool existed that understood C++ symbol mangling rules. While no commercially-available tool exists, a tool called mkCCshlib has been developed internally at Silicon Graphics for such a purpose. A version of this utility was obtained from John Wilkinson, its author, during development of the Distributed Configuration Manager. Although the mkCCshlib utility ultimately invokes the standard mkshlib utility, it shields the shared library builder from most of the difficulties involved in building a shared library. mkCCshlib automatically generates a branch table for the shared library. It also redefines any C library symbols it finds in the shared library objects. C++ library symbols must still be redefined by the shared library builder; however, mkCCshlib displays an error message for any undefined C++ library symbols with explicit directions on how to perform redefinition.

Many of these deficiencies have also been resolved in the newest version of the UNIX operating system, System V Release 4 (SVR4). SVR4 and its shared library implementation are discussed in section 4.

## 3 The DCM Process Migration System

The process migration system in the Distributed Configuration Manager consists of two components: the Migration Policy Manager and the Migration Server. The Migration Policy Manager uses user input from the Distributed Configuration Manager GUI (Graphical User Interface), translating configuration updates into migration requests for the Migration Server. This section outlines the policy used by the Migration Policy Manager and the operation of the checkpointing mechanism of the Migration Server.

## 3.1 The Migration Policy Manager

Migration is initiated on demand using the DCM graphical user interface. Two databases are used by the Policy Manager to establish policy: the Network Configuration Database and the Migration Pool Database.

The Network Configuration Database (NCD) contains a list of machines that are connected to the user's network. The Network Configuration Database is an ordinary text file and machines can be added to the network or removed from it with a text editor such as vi. Each machine entry in the Network Configuration Database contains the name of the machine, its performance characteristics such as the number of processors and clock speed, and whether the machine is available to run DCM processes. Machine availability is controlled by the DCM graphical user interface. The Network Configuration Database needs to be modified only when machines are added or removed from the network and therefore does not change often.

The Migration Pool Database (MPD) contains a list of processes that are capable of migration. This list of processes is known collectively as the migration pool and an individual entry is known as a migration pool database entry, or MPD entry. The process represented by a given migration pool entry is known as a migration candidate. Each MPD entry contains the UNIX process id (pid) of the process, its name, the full path of its location on the file system, and the processor it resides on. Like the Network Configuration Database, the Migration Pool Database is stored as a text file; however, its existence is transparent to users of the Distributed Configuration Manager. Unlike the Network Configuration Database, the Migration Pool Database changes dynamically, with process entries being added and removed from the pool during their execution.

The Distributed Configuration Manager uses the information in these two databases to determine migration policy. The current migration policy consists of the following steps:

1. The user configures the network for migration and initiates the Policy Manager using the DCM graphical user interface.

2. The Policy Manager scans the NCD for machines that have been selected as UNAVAILABLE. If an unavailable machine is found, it becomes the current source machine, as it is the source of processes that will be migrated.

3. Using the name of the source machine, the Policy Manager searches for a pool entry that represents a process executing on the source machine. This entry is deleted from the MPD.

4. Using data from the deleted pool entry, the Policy Manager makes a request to the Migration Server, which in turn invokes a utility called Slayer on the invalid machine. Slayer contains the checkpointing mechanism. It halts the process on the source machine and builds a new checkpoint file capable of being restarted on a new machine, known as the target machine.

5. Once Slayer has completed its checkpointing procedure, the Policy Manager uses the NCD to determine a valid machine to use as a target. Then, the Migration Server will restart the checkpoint file on the target machine.

6. The Policy Manager returns to step 3 until all processes are removed from the invalid source machine.

7. The Policy Manager returns to step 2 until all machines listed in the NCD have been checked for availability.

The Migration Server is currently a single-client server and is part of the Distributed Configuration Manager central process. Future versions of the Distributed Configuration Manager will contain a multiple-client Migration Server. This will allow multiple users, each running his or her own copy of the DCM graphical user interface, to request migration. A multiple-client Migration Server will also allow the connection of other types of policy managers that may provide different policies such as fault tolerance and load balancing policies. A proposed design for the multiple-client migration server is presented in section 4. Fault tolerance and load balancing are discussed in section 5.

## 3.2 The DCM Migration Mechanism - Overview

The checkpointing and restart mechanism in the Distributed Configuration Manager is an enhancement of the one found in the Condor Distributed Batch System but it offers a number of improvements. These include:

- Migration of processes that use shared libraries
- Migration of processes programmed in C++
- Upgrades to support the IRIX 4.0.5 operating system
- Increased performance by using NFS to eliminate need for file transfers
- Direct, rather than shadow, execution of system calls to reduce residual dependencies
- Object-oriented software construction in C++ to support the addition of extra capabilities

The heart of our migration mechanism is a procedure known as checkpointing. As defined by Litzkow, checkpointing is achieved by "storing the process state and later restoring it in such a way that the process can continue where it left off." [14] Much of the state of a UNIX process can be determined by its executable file and its core image. Additional state information can be obtained through the use of portable UNIX system calls. This section will describe the operation of the DCM checkpointing mechanism, which is an enhancement of the Condor checkpointing mechanism.

As mentioned earlier, the state of a UNIX process includes the contents of memory (the text, data, and stack segments), processor registers, and the status of open files. The text segment is easy to retrieve because it does not change and it can be found in the executable file. Core files are intended to aid in the program debugging process; however, the information needed to debug a process and the information needed to restart it are nearly identical. Data from the core file is copied into the new checkpoint file according to the semantics of the COFF format. The stack is also retrieved from the core file, appended to the checkpoint file, and restored during the restart process. The restart process is explained below. Restoration of the process execution state, as discussed in section 2, is difficult because recovery of information such as the hardware registers and program counter varies among hardware platforms. Fortunately, UNIX provides a generic pair of routines called setjmp() and longjmp(). These routines allow system programmers to perform non-local jumps inside of a program and are generally used to jump to an error handler. However, their operations internally store the full stack frame with a current stack pointer, the internal hardware registers, and the program counter. Because these functions are part of the UNIX standard, the burden of porting them among platforms is the responsibility the operating systems vendor. Reliance on the executable file format, core file format and setjmp/longjmp facility means that the checkpointing mechanism can be easily ported to any UNIX platform with only a few exceptions.

The migration process relies on two components. The first component is a piece of code, called the bootstrap module, that when linked allows a program to become a migration candidate. The second component is the checkpointing utility, called Slayer. Migration occurs in three steps. First, the migration candidate begins execution. The bootstrap code executes first, setting up signal handlers for checkpointing. Second, the Slayer utility sends the migration candidate a signal to begin the checkpoint sequence and creates a checkpoint once the candidate has terminated on its source machine. Third, the candidate will be restarted on the new target machine selected by the migration algorithm. Once again, the bootstrap code is responsible for re-initializing the program on the new machine. A detailed description of these three steps follows.

## 3.3 The DCM Migration Mechanism - Detailed Description

This section presents a detailed description of the DCM Migration Mechanism. In general, process migration involves three steps [1]:

1. Suspension of a process on its original (source) machine.
2. Transfer of process state to a new (target) machine.
3. Resumption of execution on the new host.

These steps are presented in the order of their operation.

### 3.3.1 Process Suspension

The bootstrap module is a set of object files that allows a program to become a migration candidate. The module contains code to add the migration candidate to the Migration Pool Database upon program startup, remove the migration candidate from the Migration Pool Database upon successful termination, and set up signal handlers to respond to the requests of the checkpointing mechanism. The bootstrap module requires the user to make no modifications to his or her code; however, the program must be linked with the bootstrap modules to become a migration candidate. Modifying the Migration Pool Database is trivial; setting up the signal handlers is not trivial and is explained here.

All programs compiled on a UNIX system are ultimately linked with a file called crt0.o or crt1.o, depending on the UNIX implementation (Silicon Graphics systems use crt1.o). The crt object file contains initialization code for the executable program. One of its functions is to inform the program linker to use the symbol main as the first procedure for execution. C and C++ programmers know that they must always write a function in their programs called main() that is always executed first. To build a migration candidate, the link step is altered to link in a replacement for the crt object file, which is called mycrt1.o in the DCM implementation. The file mycrt1.o is a modified copy of crt1.o whose only difference is that the main symbol has been replaced. The replacement symbol name is arbitrary; however, the Distributed Configuration Manager (and Condor) use MAIN for simplicity. When a migration candidate program is executed, it will execute the procedure MAIN() instead of the usual main(). The MAIN() function is included in the bootstrap module, which a migration candidate links with. In addition to performing its initialization duties, MAIN() will call the user's main() with the correct environment (argc, argv, and envp) and return the correct return value from main() upon process completion.

The initialization performed by MAIN() creates a signal handler for the TSTP signal. The TSTP signal is the terminal stop signal. Its original purpose is to recognize the process suspend key from the keyboard (usually Control-Z) that stops an executing process. Because migration has a similar purpose, a custom TSTP signal handler is used to inform the migration candidate to suspend itself for checkpointing. The TSTP handler saves the current process execution state in the process virtual address space using setjmp(). It also sets a global variable, restart, to TRUE. Then, the TSTP handler sends the process the QUIT signal, which dumps the core (required for checkpointing) and terminates the program. Once the process has terminated, the Slayer utility may be invoked to produce a new checkpoint file.

### 3.3.2 The Checkpointing Process

The DCM checkpointing mechanism is a separate process, called Slayer. Because Slayer is separate, it can be invoked directly from the command line or from inside another process using the remote shell (rsh) command, as we do. Slayer is invoked as follows:

- slayer <process pid> <source file> <checkpoint file> [core file]

The process pid parameter is the UNIX pid value that defines the specific process. The source file parameter is the name of the file which contains the text to be used in the checkpoint file. It is usually the name of the original executable file, although text can be obtained from any previous checkpoint file as well. The checkpoint file parameter is the desired name of the checkpoint file. We always name a checkpoint file <original_name>.ckpt for consistency. Finally, the core file parameter specifies the name of the core file to be used. UNIX core files are always named core, regardless of the program that created them. However, all of Slayer's file name arguments must be specified with a full path.

Slayer first sends the TSTP signal to the migration candidate, the process specified by the process pid. The migration candidate is also known as the victim. The TSTP signal activates the checkpointing mechanism in the bootstrap code of the migration candidate. The TSTP signal handler will call the setjmp() function, preserving the process execution state (register and stack frame contents) in the process virtual address space. Then, the TSTP signal handler will send the KILL signal to the process, suspending its execution on the source machine and dumping its core.

After the process has been terminated, Slayer will build a checkpoint file using the source file and core file specified. Slayer retrieves the text section and symbol tables from the source file and combines them with data from the core to create a new executable file, the checkpoint file. The original source file stored initialized and uninitialized data contiguously, but in separate sections as specified by the COFF format. The new checkpoint file is different because both the initialized and uninitialized data have been initialized according to the location of the program counter when the process was terminated. Unlike the original executable file that used the bss section header to save space, a checkpoint file has no uninitialized data. After execution, the bss section has been initialized in virtual memory to values that must be preserved for checkpointing. The bss section in a checkpoint file will contain a data in addition to its section header. For this reason, a checkpoint file is substantially larger than its original executable file.

Once the data has been retrieved from the core file, Slayer retrieves the stack data map from the core file, appends it to the end of the checkpoint file, and then appends the size of stack data map. The stack data map section is not incorporated into the COFF description; however, the bootstrap module contains code to retrieve the stack data upon restart.

### 3.3.3 Process Restart

The bootstrap code will execute again once the checkpoint file is restarted. The bootstrap determines the process location and process pid and makes the appropriate entry into the Migration Pool Database. Because the value of restart is TRUE (set previously by the TSTP handler), the bootstrap code recognizes that it is not executing for the first time and sends the process a signal that causes a custom signal handler to execute and retrieve the stack data map that was saved by the slayer application. It also retrieves the stack frame saved from setjmp(). Once these two pieces of data have been restored, longjmp() is invoked with the stack frame from setjmp(). The program will begin executing with its old stack frame and old data. At this point, the process has been successfully restarted on its new host.

## 3.4 Migration of Processes Using Shared Libraries

Our enhancement to the Condor migration mechanism allows the migration of executables that use shared libraries. A process that uses shared libraries can save system resources and increase system performance. Extending the checkpointing mechanism to support such processes adds a new and important class of processes to those supported by Condor.

Shared libraries are used to increase system performance by saving disk space, saving memory, and improving library code maintenance. Unlike a standard UNIX archive that is duplicated both on disk and in memory for each process that uses the archive, there is only a single copy of the shared library. Although the text of a shared library can be shared, each process must have a private copy of data associated with the shared library. A process migration mechanism must recover both the text and data sections of a shared library, much like the mechanisms already used to retrieve standard text and data.

As discussed in section 2, an executable file contains two additional sections when a shared library is used. The first section, lib, contains the pathnames of all shared libraries used by the program. The other section, init, contains initialization code for the shared library. When a process begins to execute, the system loader uses the information in these sections to find the shared libraries, load them into memory, and initialize them properly. This process occurs automatically when a process begins execution. Therefore, no special action is required to restore the shared library text.

When a process executes shared library text, it will execute in an identical manner to standard text; it merely exists at a different virtual address in the process address space. Global (static) variables in the shared library will exist in the reserved section (specified during the creation of the shared library) while automatic variables created in the shared library will be created on the regular process stack. These automatic variables will be restored using the current stack retrieval mechanism. However, the global data must somehow be preserved. For debugging purposes, this data is placed in the core file in a special data map called VLIBDATA along with its virtual address.

To recover this data, the bootstrap restart mechanism has been modified to extract this data from the core file and load it into the corresponding address. The restart mechanism opens the core file and searches for a VLIBDATA map, indicating that the process uses shared libraries. A VLIBDATA map and its corresponding data will exist for each shared library used by the program.

For instance, a program that uses three shared libraries will have three VLIBDATA maps in its core file. If no VLIBDATA maps are found in the core file, it can be safely assumed that the process does not use shared libraries and the restoration process is performed normally. If VLIBDATA maps are found, the map header is searched for the virtual address of the data section, its length in bytes, and its location in the core file. The data section is read from the core and copied into its correct virtual address location in the process virtual address space. Once all of the shared library data sections have been restored, the normal bootstrap mechanism continues as before.

## 3.5 Migration of Processes Interfaced to a Communication System

As mentioned in section 1.1 our long-term research goals are focused toward migrating QUEST Distributed VHDL Simulator objects.

Currently, each simulation object takes advantage of a hierarchical communication subsystem to transfer information. The communication subsystem uses shared memory for intramachine communication and distributed shared memory (SCRAMNet) or Ethernet for intermachine communication. Each simulation object maps local shared memory into their virtual address space and does not know about SCRAMNet or Ethernet.

In order to migrate the QUEST communicating objects mentioned in the previous paragraph, the migration facility must be able to unmap a local shared memory area prior to the checkpoint, and then map in a new shared area after being relocated on a target machine. The shared memory space can not be treated as part of the process state and migrated. This is because each shared memory space exists, and is managed independently from shared memory spaces on other machines. This task requires support from the communication subsystem. The following subsections describe the steps taken to migrate a communicating process.

### 3.5.1 Process Suspension

The TSTP signal handler (described in section 3.3.1) was modified to call a migrate() routine prior to saving the process execution state with the setjmp() call.

The migrate() routine is a communication primitive that places a hold on any existing messages en-route to the process to be migrated. After the delivery process has been put on hold, the process de-registers itself as a participant in the shared memory communication system and unmaps the shared memory. Once the communication space is unmapped, the process can proceed to the checkpointing process.

This scenario is similar to the process of a person who is moving. Before moving, mail is put on hold (actually forwarded to somewhere new). The living space is eventually relinquished (house sold) and the person leaves the local neighborhood.

### 3.5.2 The Checkpointing Process

There is no special action required to checkpoint a communicating object.

### 3.5.3 Process Restart

When the Distributed Configuration Manager chooses a target machine, the object identifier is supplied to the communication system manager on the target machine. This allows the manager to allocate a spot in the target machines registration area for the migrating object.

When a migrating process is restarted, the object registers itself with the taget machines communication system manager and maps the new local shared memory into its virtual address space. This task is performed prior to issuing the longjmp() call described in section 3.3.3.

Finishing the moving person analogy from section 3.5.1, this is similar to finding a place to live (allocate a spot), moving in (register with the communication system), and continuing with life (issue the longjmp() call).

### 3.5.4 Dealing with Communication System Managers

In addition to migrating objects interfaced to the communication system, the ability to alter the configuration (list of available/usable machines) of the communication system itself is important. In other words, if a machine will no longer be available for a VHDL simulation, then all objects must be migrated and the communication manager should be removed.

In such a scenario, there is no target machine for the communication manager to migrate to. The reason for this is that the communication manager was set up specifically to handle communications on the source machine which is now to be evacuated.

The communication manager uses sockets (Ethernet) and SCRAMNet to enable intermachine communication. If it were simply terminated, the processes at the other end of the sockets would be left in unpredictable states, and may themselves hang.

Since the migration utility currently does not support migration of processes with sockets, the communication manager can not be checkpointed and left in a migration state until the machine becomes available again or the simulation terminates. Ankola's work at the University of Cincinnati [1] allows for the migration of processes using sockets. However, his work does not maintain the original socket connections. Instead, the original sockets are closed and new sockets are opened to take their place.

Currently, this is an outstanding problem which must be addressed if the Distributed Configuration Manager is to be used effectively. A solution for shutting the Ethernet sockets down has been developed, but not fully implemented.

## 3.6 Summary

This section described the design and implementation of an enhanced process migration facility for UNIX workstations. The migration mechanism has been modified to allow the migration of processes that use UNIX shared libraries. Unlike most process migration mechanisms, the mechanism in DCM requires no modifications to the kernel and has been successfully demonstrated on a network of Silicon Graphics 4D workstations. Processes may freely migrate without restriction or dependency from single-processor desktop workstations to high-end multiprocessor machines. Processes interfaced to the VHDL simulation communication system can be migrated, however dealing with the communication managers remains an unresolved issue. The next section evaluates the DCM system and presents some design alternatives and enhancements for future versions.

# 4 Results / Evaluation of Work

Having described our implementation of a process migration mechanism of the Distributed Configuration Manager in the previous section, attention is now focused on analyzing the qualities of the finished system. This section analyzes the Distributed Configuration Manager migration system with respect to the four original design considerations discussed in section 2: transparency, system interference, residual dependencies and complexity. Performance and software construction are also considered. Some enhancements to the system are also proposed. Finally, the limitations of our system are addressed. The chief weakness of the DCM process migration mechanism is its vulnerability to operating system upgrades. A minor upgrade to the operating system during development posed a substantial challenge and future upgrades to the operating system will pose an even greater challenge. However, our modularized version is much less complex than original Condor. Some strategies for dealing with this weakness are presented shortly.

## 4.1 Achievement of Design Goals

In designing the process migration mechanism, four main design considerations were addressed, defined in section 2. These goals were transparency, minimal interference with the system, residual dependencies, and complexity. In this section, the implementation of migration in the Distributed Configuration Manager is analyzed with respect to these design considerations.

### 4.1.1 Transparency

A process migration system should have a network-transparent execution environment. The DCM system achieves transparency through the use of the Network File System, or NFS. NFS provides each workstation on the network with an identical view of the file system and a consistent global naming scheme. It is possible to migrate a DCM pool process to any workstation on the network without restriction.

### 4.1.2 System Interference

Process migration should not introduce excess interference with either the process being migrated or the system as a whole. The migration mechanism should operate in an atomic fashion with respect to the system. This consideration becomes especially important in cases such as interprocess communication where a time-out failure may result if the time between process suspension and restart is too great. Our present system does not interfere with the processes being migrated or the migration system. Processes can be halted, checkpointed and restarted at will. A checkpointed process is not required to restart immediately after checkpointing; it may reside on the file system and be restarted at will.

### 4.1.3 Residual Dependencies

The process migration mechanism should be designed to minimize residual dependencies on previous locations. For instance, once a process has been migrated from Machine A to Machine B, the process should no longer require Machine A. This is an important design principle especially when implementing a fault tolerance mechanism. If a process is moved from Machine A to Machine B but the process still has residual dependencies on Machine A, the process will crash if Machine A fails. We seek to avoid such limitations.

Residual dependencies have been minimized but not completely eliminated. Unlike the Condor system, a shadow call mechanism has not been implemented. Such a mechanism was desirable in Condor because not all systems on which Condor must run have a network-transparent file system such as NFS. The shadow call mechanism imposed a residual dependency on the host where shadow calls were executed. While we have eliminated the shadow call dependency, each process in the Distributed Configuration Manager migration pool has a residual dependency on the

Distributed Configuration Manager main process. Checkpointed files are restarted from the main process using the Berkeley remote shell (rsh) command. Because of the semantics of the rsh command, a parent-child relationship is enforced between the Distributed Configuration Manager and a restarted application. If the Distributed Configuration Manager abnormally terminates during the execution of checkpoint files, the remote shell system will fail resulting in the premature termination of all checkpoint files. This dependency is due to the semantics of the rsh command and is unavoidable.

### 4.1.4 Complexity

In a custom operating system such as Sprite, complexity is an important factor to consider and minimize. Process migration affects virtually every major piece of an operating system kernel. A maintainable migration mechanism must limit its impact on the operating system [7]. Even though migration has been implemented outside of the kernel, complexity must still be reduced. The migration mechanism used by the Distributed Configuration Manager is a large, complex piece of code. Each feature of the UNIX operating system that is supported, such as shared libraries or communication, requires an additional subsystem to the standard migration mechanism.

The migration system is additionally complex because it relies on UNIX mechanisms that are subject to change during operating system upgrades. The designers of the systems such as Sprite and Charlotte do not have this problem. They have total control of design changes and can make choices designed to minimize the overall impact of these changes. In systems such as the Distributed Configuration Manager, changes to the operating system are completely beyond our control. Minor changes in the UNIX operating system can result in substantial development time spent on patching the migration mechanism. This insensitivity to change minimizes the portability of our system, and therefore increases complexity. The issue of portability is further discussed in section 4.4.

### 4.1.5 Performance Measurements and Verification

Although obtaining peak performance is important, it has never been an overriding issue in the design of our system. Convenience and resource management have been the key motivating factors. The Distributed Configuration Manager has been verified using test programs. These test programs include simple programs to perform mathematical computations. We used the UNIX sleep() command to temporarily halt the execution of our test programs between computation of a value and the display of its result. The sleep command provided us with a time window within which we invoked Slayer from the command line. The resulting checkpoint file was restarted on a new host, where the displayed computational value could be compared to its intended result. We also constructed programs that tested dynamic memory allocation in a similar manner. Once we built our recovery mechanism for shared libraries, we constructed a shared library using mkCCshlib and used the shared library in the construction of additional test programs. Our shared library contained routines from our original test programs. We used an assortment of local and global data, providing test cases for the recovery both automatic and static variables. We again tested the dynamic allocation of memory from within our test programs.

Performance data has been obtained in the Condor [10] and apE [1] systems. While shutdown and startup costs are relatively minor, the greatest performance hit occurs during the checkpointing process. The time required to checkpoint increases in a roughly linear fashion with the size of the virtual address space. The existence of NFS at our execution site allows us to eliminate the Condor shadow mechanism and the need to copy checkpoint files from machine to machine. However, NFS imposes its own overhead and makes our system sensitive to the network traffic of other users.

Process startup and shut down occurs in a few milliseconds. Checkpointing varies from two to ten seconds depending upon the size of the files involved and network activity on the file system. A typical restart occurs in a few milliseconds; however, Ankola discovered that requesting UNIX communication paths added up to an additional ten seconds to his restart mechanism. Process migration is an expensive process, feasible only for long-running programs such as simulations. Short jobs do not benefit from using a process migration system.

## 4.2 Migrating C++ Programs

One of the limitations of Condor is that it supports only applications programmed in C and FORTRAN. Our experience has shown that the migration of processes created with C++ requires some additional steps. These steps are outlined here.

A nice implementation feature of Condor is that it does not require its users to recompile their applications. This mechanism provides transparency for our users and it was our intention to preserve this convenience for the Distributed Configuration Manager. Condor does, however, require users to link their code with special bootstrap routines that include the startup and shutdown routines necessary for the migration sequence.

As we explained in the previous section, recompilation is avoided by replacing the system crt1.o file with a modified version, in our case, mycrt1.o. The code in crt1.o contains initialization code that gets executed upon the startup of a program. It contains the definition of the first symbol to be executed after startup. The C language requires the programmer to write a procedure entitled main() that defines the starting point of the program. The routines in crt1.o therefore define main() as the starting symbol. By replacing crt1.o with our modified version, we can select a new symbol name for startup. For simplicity, we chose MAIN() as our starting symbol. The bootstrap code, linked into all migration candidates, begins with a procedure called MAIN(). This function is responsible for setting up migration signal handlers and calling the main() of the original application. In this manner, we have transparently set up startup and shutdown routines for the migration candidate without modifying code.

This system works well for C programs; however, when we tried to use the same format for C++ programs, it failed. After much trial and error, we determined that our programs failed at the location of the first C++ library function in our program, but we were at first unsure why. It was obvious that the error was occurring due to our special linking procedure, so we compiled a test program with the C++ using its verbose option to list all of the steps involved in compilation. While we had used the verbose (-v) option of the C compiler to view the link sequence, the C++ preprocessor contains an additional option (+v) for verbose output. When we used this additional option, we discovered a little-known step of building a C++ program. Most C++ compilers consist of a front end, often called cfront, that converts C++ into C code, which is then compiled by the standard C compiler. However, when the C compiler is complete, a program called c++patch is executed, using the name of our program as an argument. Most of the classes in the system C++ library contain global constructors that must execute prior to the execution of user code. The c++patch utility modifies the executable file to insure the correct initialization of these constructors.

We discovered that calling c++patch was not the only problem affecting the execution of programs using C++ code. The global constructors are initialized in the main() function, which does not execute until after our bootstrap code. Our original bootstrap code used C++ library functions such as cout. Attempting to execute our bootstrap resulted in additional core dumps. We chose to rewrite the bootstrap code using only C library calls, such as printf(), eliminating the problem. We could have chosen to execute the C++ global constructors from our bootstrap code; however, it is possible that the C++ libraries could change their implementation, making this approach invalid. Restricting bootstrap programming to only C library calls is much easier. This constraint on the content of bootstrap code should be observed if future modifications to the Distributed Configuration Manager require modifications to the bootstrap module.

## 4.3 Suggested Enhancements to DCM

This section proposes several enhancements to the current DCM implementation. Future research directions are presented in the next section; the enhancements presented here are aimed at providing greater ease-of-use, better portability and reduced program maintenance.

### 4.3.1 Suggested Enhancements to the Distributed Configuration Manager GUI

The current GUI provides an easy-to-use front end to control network configuration and migration policy. The GUI has been implemented using the GL language. GL is suited for visualization and animation, and not particularly robust for event-driven user interfaces. GL lacks libraries for the creation of user interface constructs such as buttons, pull-down menus and dialog boxes. A good windowing environment provides libraries for these objects and also performs automatic detection of events such as window resizing and the pressing of pointer buttons. Our GL interface contained a main loop that checked for these events and took corresponding action. Such code would have been unnecessary in a true windowing environment. A fairly simple enhancement would be replacing the GUI with a MOTIF-based GUI. MOTIF is currently the de-facto standard for UNIX workstations. It provides libraries for user interface objects such as push buttons, creates prettier screens, and maintains a consistent look-and-feel with other UNIX applications. MOTIF will also increase the portability of our user interface, an issue discussed further in Section 4.4.

Another suggested enhancement is the addition of additional menus to allow graphical modification of the Network Configuration Database. This capability would allow a user to add machines or groups to the database without directly editing Network Configuration Database files or understanding their file format.

### 4.3.2 Suggested Enhancements to the DCM Policy Manager

The DCM Policy Manager consists of the Network Configuration Database (NCD) and the Migration Pool Database (MPD). This data model is implemented in a manner that allows only a single client (the DCM Configuration Policy Manager) to access it. A better implementation would separate the location of the databases from the Policy Manager, providing multi-user capability. This would permit simultaneous copies of the DCM GUI and Policy Manager to execute. If the databases are modified to permit multiple DCM policy managers, they will easily support access by other kinds of policy managers. Our policy is an on-demand policy that uses process migration for network configuration. Multi-client databases could allow the connection of additional policy managers that implement fault-tolerance and load balancing policies. Although the database problems, such as consistency and serialized access, associated with multiple clients are well known, we suggest the use of a commercially-available database library or database system to address these problems and minimize development time.

### 4.3.3 Suggested Enhancements to the DCM Migration Server

The current migration server is only capable of migrating one process at a time. Some performance gains can be achieved through the parallelization of migration requests. The migration server should be modified to run as a stand-alone daemon, capable of accepting simultaneous requests from multiple policy managers. A stand-alone migration server would be capable of parallelizing migration requests. However, performance gains obtained through parallelization may be diminished by the increased NFS traffic generated during checkpointing.

## 4.4 Portability Issues and OS Upgrades

Software maintenance is becoming an increasingly important issue in the software development world [24]. Software maintenance is required when bugs are discovered in an existing system, or when software is ported from one platform to another. A piece of software is said to be portable if it can be moved from one platform to another, ideally with little or no modification. Software may be ported between differing hardware platforms, such as from Silicon Graphics 4D systems to Sun Sparc systems, or it may be ported between different software platforms. A software platform changes when an entirely new operating system is installed. One example for an upgrade an IBM PC from MS-DOS to a UNIX variant such as NeXTSTEP. A more subtle, yet insidious, software platform change occurs when a hardware vendor releases an upgrade to an existing operating system.

The usefulness of a software application is greatly enhanced if the application is highly portable. A software application that is supported on several hardware platforms is marketable to a larger set of users than an application supported on a single platform. Likewise, the usefulness of a software application is enhanced if it continues to run with only minor modifications as the operating system is upgraded. Portability is generally a market concern; however, because the Distributed Configuration Manager is a usable piece of application software in addition to being a research platform it is important to strive for ease of software maintenance wherever possible in our design. This section analyzes the degree of portability of the DCM application with respect to both hardware and software.

### 4.4.1 Portability of the DCM Graphical User Interface

The DCM Graphical User Interface has been designed and developed using the GL language, a graphics description language developed by Silicon Graphics. GL has traditionally been available only on Silicon Graphics systems, mainly as a marketing feature. However, the software marketplace as a whole is embracing an idea known as open systems. A system is said to be an open system if it is widely available and capable of connecting and functioning together with different kinds of systems. The opposite of an open system is a proprietary system. The UNIX operating system is generally acknowledged as being an open system due to its wide availability on multiple platforms, whereas the MS-DOS operating system is not. Not wanting its products to be labeled as proprietary, Silicon Graphics has developed a version of the GL language known as OpenGL. OpenGL can be installed on a wide variety of operating systems, including SunOS.

Although the GL language is becoming a more open standard, there are other factors that limit its use in our system. GL is a graphics description language and tends to be well suited for rendering and animation; however, it is not intended to be used for user interface design. Rendering and animation are process-driven activities; requesting input from a user is an event-driven activity. GL programming was used in an existing prototype of the Distributed Configuration Manager interface and it was modified, rather than rewritten, for the existing Distributed Configuration Manager application.

Future versions of the Distributed Configuration Manager should base their interfaces on a GUI standard, such as the MOTIF windowing interface. MOTIF is the de-facto standard windowing interface for UNIX workstations. MOTIF has even been embraced by Sun Microsystems, which is abandoning its proprietary OpenLook window interface. Porting a MOTIF-based interface for the Distributed Configuration Manager would be as simple as recompiling; no modifications would be required to the code.

Much design work has already been performed in anticipation of a future redesign of the user interface. Through careful use of object-oriented design, the data structures that control the GUI have been removed from the data structures that control program behavior. These data structures have formed the basis for C++ classes. The DCM graphical client consists of a GUI class, while the program behavior, particularly the Migration Policy Manager and its two databases, have been organized into a separate class. There is a minimal set of well-defined dependencies between these two classes. A MOTIF port would merely require redesigning the existing user-interface class and replacing the existing class. The class design of the Distributed Configuration Manager GUI is a good example of using objects as building blocks. The implementation of an object in a software system should be able to be replaced without affecting the operation of other objects in the system. We have achieved this degree of modularity in the Distributed Configuration Manager GUI.

### 4.4.2 Portability of Process Migration Mechanism

The process migration mechanism used in the Distributed Configuration Manager should be quite portable amongst various UNIX platforms. Because the mechanism relies on the COFF format and C library functions such as setjmp() and longjmp() as well as generic file operations, most of the porting work for the Distributed Configuration Manager has already been performed by the designers of the operating system. The Condor system, upon which our process migration mechanism is based, has been ported to over ten different UNIX platforms.

In practice, however, portability is not as straightforward; there are still enough variances between UNIX systems to make each port a large job. A close examination of the Condor source code reveals that a completely different checkpointing mechanism exists for each UNIX variant supported. From a software maintenance point of view, we feel that object-orientation greatly simplifies the multi-platform design of Condor. Object-orientation allows a single class hierarchy to be developed and maintained for the checkpointing mechanism. A base class is designed to implement a generic checkpointing mechanism and derived classes are designed for specific architectures. This allows common procedures and data structures to be used by all Condor variations and additionally allows redefinition of unique procedures through the use of function overloading, or as it is known in C++, virtual functions.

We cannot emphasized enough that the use of object-orientation will not reduce the complexity of this problem; it only helps to manage the existing complexity. Fortunately, there may be a silver lining to our portability cloud. A growing movement towards standardization in the UNIX world seeks to eliminate many of the minor differences between various implementations. Only time will tell whether the rhetoric of the standards bodies will materialize into a unified UNIX standard. At this writing, most of the UNIX systems vendors are releasing newer versions of their operating systems based on the System V Release 4 (SVR4) Applications Binary Interface (ABI) [22]. Although the goal is enhanced portability, SVR4 poses many problems for the Distributed Configuration Manager. These problems will be discussed in Section 4.4.4.

### 4.4.3 Porting the Process Migration Mechanism from IRIX 4.0.1 to IRIX 4.0.5

The original Condor checkpointing mechanism made assumptions about the ordering of the data sections in the SGI extended COFF format. These assumptions were true for previous versions of IRIX such as IRIX 3.3.1 (used by the developers of Condor) and IRIX 4.0.1, which was the current version at the time we were porting the checkpointing mechanism. However, the ordering of the data sections was modified for the maintenance upgrade IRIX 4.0.5. It is difficult to know why the data sections were reordered, especially in light of a misleading comment found in the SGI header file <a.out.h>:

"Coff files produced by the mips loader are guaranteed [emphasis added] to have the raw data for the sections follow the headers in this order: .text, .rdata, .data and .sdata the sum of the sizes of last three is the value in dsize in the optional header."

Of course, we brought the upgrade problem upon ourselves by relying on the system documentation. Relying on system documentation is sheer folly and will always result in disaster. A comment in a system header file is not necessarily correct just because the vendor went to all the trouble of placing it there. When our development systems were upgraded to IRIX 4.0.5, purported to be a minor bug fix, we discovered that the ordering of the rdata and data sections had been reversed. Once this reversal was discovered, our checkpointing algorithm was modified to ignore proper section names and rely instead on the contiguous virtual address space of the data area. The sections in the data area are always contiguous in the process virtual address space. Our new solution proved robust enough to work with both the old and new versions of the operating system, and hopefully we expect it to continue to work with future versions of IRIX (more about that in the next section).

Although this upgrade appears trivial, it required three weeks to diagnose and correct. Three weeks is a significant amount of time, especially when one considers that the upgrade was supposed to be a minor bug fix.

### 4.4.4 Porting to Future Versions of IRIX

The current version of the Distributed Configuration Manager was written under IRIX 4.0.5. Significant changes will be introduced when Silicon Graphics releases IRIX 5 to its customers. Release 5.0.1 is currently shipping on the new Onyx/Challenge family and existing systems will support IRIX 5 with release 5.1, tentatively scheduled for sometime in fourth quarter 1993. IRIX 5 is a System V Release 4 implementation; previous versions of the operating system were based on System V Release 3. While a majority of these changes will be transparent to the casual system user, they have dire consequences for applications, such as the Distributed Configuration Manager, that rely on underlying UNIX mechanisms. Significant changes will have to be made to the Distributed Configuration Manager to support IRIX 5. The upgrade from IRIX 4.0.1 to IRIX 4.0.5 required only minor changes to existing code; however, support of IRIX 5 and its underlying SVR4 implementation will require the process migration mechanism to be completely rewritten.

Problems with operating system changes is the Achilles Heel of the Distributed Configuration Manager and the Condor Distributed Batch System. Although typical applications developers do not care how the UNIX system builds and executes programs, we are intimately concerned. We do not access internal kernel structures; however, we make full use of knowledge concerning file formats such as COFF and the operation of the program loader; these are parts of the UNIX OS whose implementation is subject to change.

While the cost of software maintenance is high for a system such as DCM, it is a price that must be paid to achieve our design goals. This topic is further discussed in the next section.

## 4.5 UNIX System V Release 4

There are three major feature changes in System V Release 4 that differ from previous versions of the operating system. These are: The Executable and Linking Format (ELF), Dynamic Shared Objects (DSO) and Position-Independent Code (PIC). The ELF format replaces the COFF format used in previous versions of the operating system and Dynamic Shared Objects replace the current static shared library implementation. The following sections detail these new features and contrast them to those found in existing implementations.

### 4.5.1 COFF vs. ELF

Versions of System V prior to Release 4 (including IRIX) used an extended version of the Common Object File Format, or COFF, for object files. ELF (Executable and Linking Format) is a new format specified by the System V Release 4 Applications Binary Interface (also known as the SVR4 ABI). Although the SVR4 loader will continue to execute COFF binaries created under previous versions of the operating system, the SVR4 compiler system produces ELF object files instead. ELF objects and COFF objects may not be mixed, meaning that if any modules of an existing application program are recompiled under SVR4, the entire program will have to be recompiled as well. Additionally, ELF provides support for dynamic shared objects, which are more flexible than the shared library implementation in SVR3. There are three kinds of ELF object files:

- Relocatable files contain code and data in a format suitable for linking with other object files to make a shared object or executable. Intermediate object files created by the compiler are examples of these.

- Dynamic Shared Objects contain code and data suitable for dynamic linking. Relocatable files may be linked with dynamic shared objects to create a dynamic executable. At runtime, the runtime linker combines the executable and dynamic shared objects to produce a process image.

- Executable files are programs ready for execution. These may or may not use the dynamic linking feature.

According to the IRIX System Programming Guide, IRIX 5.0 will execute all binaries which are compliant with the SVR4 ABI, as specified in the System V Applications Binary Interface-Revised Edition and the System V ABI MIPS Processor Supplement; however, binaries compiled under IRIX 5.0 may not necessarily comply with the SVR4 ABI. This means that the designers of IRIX may waver from the generic SVR4 specification, reducing the portability of any standard mechanism that may be devised for SVR4 implementations.

### 4.5.2 Static vs. Dynamic Shared Libraries

The current shared library implementation under SVR3 uses what are known as static shared libraries. Although the library code is not bound to a specific process until run time, the virtual address of a static shared library exists in a process' virtual address space is statically bound during the compilation and linking procedure.

Dynamic shared objects replace the functionality of the static shared libraries and offer additional benefits. The principal advantages of dynamic shared objects are that they do not need to be created with fixed load addresses, unlike static shared libraries, and that they can be dynamically loaded under programmatic control. Both of these features are possible because of the new runtime linker, rld, which can resolve external references between objects and relocate objects at runtime. Dynamic shared objects avoid the potential address conflicts that multiple static shared libraries may present.

### 4.6 Summary

This report has demonstrated that migration of processes that use shared libraries is possible. We have additionally built a system that provides an on-demand migration policy that can be easily extended to support additional policies, such as fault-tolerance and load balancing. However, our system is extremely sensitive to changes in the UNIX operating system. In particular, many UNIX platforms will be supporting the System V Release 4 standard, which contains many changes not present in the current version (System V Release 3). While further study of the System V Release 4 is necessary, we are able to conclude that the software cost of modifying the Distributed Configuration Manager to support SVR4 is high. The ELF format has its roots in the COFF format; however, its methods of storing data are different than its predecessor. Most of the changes in ELF have been made to support dynamic objects. Dynamic objects differ radically from the current static shared library implementation.

## 5 Conclusions/Future Research Opportunities

### 5.1 Results

This report has demonstrated the successful design and implementation of a process migration system that provides network configuration capabilities to users of a UNIX network. Our main contribution was extending an existing process migration mechanism from the Condor Distributed Batch System to support the migration of processes that use UNIX shared libraries. Shared libraries have been demonstrated to reduce code size and improve system performance. Extending the Condor migration mechanism to migrate such processes is beneficial, especially considering that the typical user of a process migration system is running large, computation-intensive programs.

In addition to extending the process migration mechanism, we have created an object-oriented framework designed to support additional research in the area of process migration. The current version of the Distributed Configuration Manager uses a policy of network configuration to drive the migration system. However, future research can be aimed at extending the existing system to support policies of load balancing and fault tolerance. With the ultimate goal of providing a migration system for the QUEST Distributed VHDL Simulator, proper handling of communication managers needs to be implemented when a machine is to be removed from the available list.

### 5.2 Conclusions

While a process migration system may be constructed, we are unsure of its true benefits. Such benefits can not be realized until our application is in general use. Because our system operates outside of an existing operating system kernel, it is very sensitive to even minor changes in operating system operation. We have modularized the system dependent portions of our code to minimize this sensitivity, but such sensitivity still exists. In addition, performance of process migration is limited and not feasible for small computational jobs. However, performance limitations are minimized because the distributed simulation objects considered for migration are sufficiently large. Therefore, migration will consist of only a small fraction of total execution time.

The state of hardware performance continues to advance; however, software performance requirements tend to always outpace any current level of hardware performance and we expect this trend to continue. Systems such as the Distributed Configuration Manager will always be useful to high-performance computer users.

## 5.3 Future Research Opportunities

Our process migration system is currently controlled by a simple network configuration policy; however, process migration is a mechanism upon which more sophisticated policies may be constructed. Of particular interest in the research community are systems that promote fault tolerance and dynamic load balancing.

Fault tolerance is the capability of insulating process execution from hardware crashes. Using the shutdown and startup mechanisms developed in the Distributed Configuration Manager, we may construct policies that allow users to completely halt processes when notified of an impending fault and they can continue execution after the fault has been eliminated.

In a distributed system, Some machines in the system are usually more heavily loaded than others. Using our process migration mechanism, we can build a system that migrates processes from heavily loaded machines to lightly loaded machines to evenly distributed workload and provide even system performance. Although such a system is conceivable, it must be constructed and tested to determine what kinds of performance gains are possible.

## References

[1] M. Ankola, *Implementation of Process Migration in apE* Master's Thesis, University of Cincinnati, 1992.

[2] Y. Artsy and R. Finkel, *Designing a Process Migration Facility: The Charlotte Experience* IEEE Computer, pp. 47-56, September 1989.

[3] A. Brickner, M. Litzkow, M. Livney, *Condor Technical Summary, Version 4.1b* University of Wisconsin, October 1991.

[4] Grady Booch, Object-Oriented Design with Applications, The Benjamin/Cummings Publishing Company, Inc., ISBN 0-8053-0091-0, 1991.

[5] D. Charley, T. McBrayer, D. Hensgen, P. A. Wilsey, M. Ankola, *Distributed Simulation on a Reconfigurable Network Using Non-Uniform Message Passing*, Proceedings of the Fifth ISMM Parallel and Distributed Computing Systems Conference, 1992. Also presented at VHDL International Users Forum Fall 1992 Conference and Exhibition, October 1992.

[6] P. Chawla, H. W. Carter, P. A. Wilsey, *An Investigation of the Performance of a Distributed Functional Distributed Simulator* 32nd Midwest Symposium on Circuits and Systems, pp. 470-473, 1989.

[7] F. Douglis and J. Ousterhout, *Transparent Process Migration: Design Alternatives and the Sprite Implementation*, Software- Practice and Experience, 21(8), pp. 757-785, August 1991.

[8] Gintaras R. Gircys, Understanding and Using COFF, O'Reilly and Associates, Inc.,ISBN 0-937175-31-5, 1988.

[9] M. Litzkow, M. Livny and M. W. Mutka, *Condor- Hunter of Idle Workstations* Proceedings of the Eighth International Conference on Distributed Computing Systems, 1988.

[10] M. Litzkow and M. Livny, *Experience with the Condor Distributed Batch System* Proceedings of the IEEE Workshop on Experimental Distributed Systems, 1988.

[11] M. Litzkow, *Remote Unix Turing Idle Workstations Into Cycle Servers* Proceedings of the Summer 1987 USENIX Conference, 1987.

[12] M. Litzkow, *Condor Installation Guide* University of Wisconsin, Madison, WI, September 1991.

[13] M. Litzkow, Response to electronic mail correspondance, July 20, 19 93.

[14] M. Litzkow and M. Solomon, *Supporting Checkpointing and Process Migration Outside the Unix Kernel* Proceedings of the 1992 Winter USENIX Conference, 1992.

[15] T. McBrayer, D. Charley, P.A. Wilsey, D.A. Hensgen, *A Parallel Optimistically Synchronized VHDL Simulator Executing on a Network of Workstations* VHDL International Users Forum Fall 1992 Conference and Exhibition, October 1992.

[16] J. Myers, *Project Update: Design of an Airborne Graphics Generator* VHDL International Users Forum Fall 1992 Conference and Exhibition, October 1992.

[17] Silicon Graphics, Inc., Assembly Language Programmer's Guide, SGI Document Number 007-0730-030, 1991.

[18] Silicon Graphics, Inc., IRIX Programming Guide Volume II, SGI Document Number 007-1440-010, 1991.

[19] Silicon Graphics, Inc., IRIX System Programming Guide, SGI Document Number 007-1794-010, 1993.

[20] J. M. Smith, *A Survey of Process Migration Mechanisms* Operating Systems Review, 22(3), pp. 28-40, July 1988.

[21] M. Theimer, K. Lantz, and D. Cheriton, *Preemptable Remote Execution Facilities for the V-System* Proceedings of the 10th Symposium on Operating System Principles, pp. 2-12, December 1985.

[22] Unix Systems Laboratories, System V Release 4 Applications Binary Interface- Revised Edition, UnixPress/Prentice-Hall, ISBN 0-13-889410-9, 1992.

[23] Unix Systems Laboratories, System V ABI MIPS Processor Supplement, Unix Press/Prentice-Hall,ISBN 0-13-880170-3, 1992.

[24] Edward Yourdon, Decline and Fall of the American Programmer, Prentice-Hall, ISBN 0-13-203670-3, 1992.

# INVESTIGATION OF THIRD ORDER NONLINEAR OPTICAL AND ELECTRO-OPTIC PROPERTIES OF STRAINED LAYER SEMICONDUCTORS WITH APPLICATION TO OPTICAL WAVEGUIDES

M. J. POTASEK
Research Professor
Department of Applied Physics

Columbia University
New York, New York 10027

Final Report For:
Summer Research Extension Program
Wright-Patterson

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington,DC
and
Columbia University
New York, New York

December 1993

# INVESTIGATION OF THIRD ORDER NONLINEAR OPTICAL AND ELECTRO-OPTIC PROPERTIES OF STRAINED LAYER SEMICONDUCTORS WITH APPLICATION TO OPTICAL WAVEGUIDES

M. J. Potasek
Research Professor
Department of Applied Physics
Columbia University

## Abstract

This research investigated the properties of strained layer III-V semiconductors with particular emphasis on nonlinear optical applications. In order to broaden the scope of the research both third order nonlinearities and electro-optic effects were considered.

# INVESTIGATION OF THIRD ORDER NONLINEAR OPTICAL AND ELECTRO-OPTIC PROPERTIES OF STRAINED LAYER SEMICONDUCTORS WITH APPLICATION TO OPTICAL WAVEGUIDES

M. J. Potasek

## Introduction

The prospect of integration of several components such as detectors, lasers and modulators make semiconductor materials advantageous for applications. In particular, the III-V semiconductors are of considerable interest for applications in the visible and near infrared spectral regions. In order to guide the light waveguides are often used. Semiconductor waveguides are of interest for applications ranging from communications to computing.

Recent research efforts have focussed on the quantum confined Stark effect. However because of the electron hole recombination dynamics, devices based on this concept are rather slow. As a result, all optical processes are being actively pursued. These advances use the third order nonlinear susceptibility which encompasses the intensity dependent change in the refractive index (n2 is the coefficient of this nonlinear index). In intensity dependent media, the phase of the wave changes as a function of distance giving rise to phenomena which are used for diverse applications. These have included the two channel nonlinear directional coupler (1-3), nonlinear Mach-Zehnder interferometer (4-5), polarization switches in birefringent fibers (6-7), two core fiber nonlinear directional couplers (8-9), and semiconductor multiple quantum well (MQW) waveguides (10-11).

However, the study of third order nonlinearities often require fast pulsed lasers. Therefore in order to broaden this investigation and incorporate features which make use of more near term laser sources, we have also investigated the use of coupled semiconductor travelling-wave amplifiers. This system is significant because it does not involve the Quantum confined Stark effect and yet is electro optic in nature which blends the applications with other electro optic devices. In this case, a single semiconductor amplifier is fabricated much like a semiconductor laser with the exception of cleaved ends on the waveguide.

In the beginning sections we will describe research on the third order nonlinearity and in the later sections we will discuss the electro optic travelling wave systems.

In general the wave equations govern the waveguide phenomena, as described below

$$\nabla\times\nabla\times E + \frac{1}{\varepsilon_o c^2}\frac{\partial^2 D}{\partial t^2} = 0 \qquad (1)$$

where E is the electromagnetic field and D is the displacement vector. The displacement vector is expanded in terms of E as

$$D = \varepsilon_o \chi * E$$

$$= \varepsilon_o \int_{-\infty}^{t} dt_1\, \chi^{(0)}(t - t_1) E(t_1) \qquad (2)$$

$$+ \varepsilon_o \int dt_1 \int dt_2 \int dt_3\, \chi^{(3)}(t - t_1) E(t_1) E(t_2) E(t_3)$$

For the slowly varying envelope of the electromagnetic field (q), the light propagation is governed by

$$iq_z + \tfrac{1}{2}\beta_2 q_{tt} + \delta|q|^2 q = 0 \qquad (3)$$

where delta includes the intensity dependent index of refraction.

## Material Properties

As stated previously, the materials of interest are the III-V semiconductors; in particular, materials such as InGaAs/GaAs and InGaAs/AlGaAs. These are strained layer materials which offer band gap engineering flexibility often not attainable in lattice matched materials. Significant parameters including the well/barrier combination, the level of strain and the material concentration in layers can be varied to modify the physical parameters such as the band-gap, electro optic and optical properties. Additionally these materials have band gaps in the technologically important regions of 0.85 um to 1.5 um. Furthermore, recent computations indicate that these compounds may have large optical nonlinearities.

In the case of thin layers, the layer mismatch induces internal strain rather than dislocations (12,13). The presence of the built in strain affects the structural aspects of the materials and influences their electronic properties through strain induced changes of the band structure. Enhanced optical properties are predicted (14-15). As an example of the strained superlattice, we cite the InGaAs/GaAs materials in which the InGaAs is the well material and the GaAs is the barrier material. In this case, the smaller bandgap material is under compression and the larger bandgap material is unstrained. For small amounts of Indium in the material, the barrier heights are relatively low which can make the supperlattice effects more prominent (16-18).

For strained layer superlattices with [111] growth axis, the orientation of the lattice constant mismatch induced strains result in polarization fields directed along the growth axis. Group III-V semiconductors are piezoelectric and strains can lead to electric polarization fields. Because one of the materials is in biaxial tension and the other is in biaxial compression, the electric polarization vectors are of opposite sign. The internally generated electric fields modify the electronic properties of the superlattice,while the internal fields are screened by photogenerated carriers which leads to large optical nonlinearities. The optical matrix elements are altered because both the electronic energy levels and wavefunctions are altered by the internal fields. In InGaAs/GaAs the fields shift the conduction band state to lower energy and the valence band to higher energy. This effect reduces the band gap. Since this is a type 1 superlattice the smaller band gap Ga alloy is in biaxial compression and the light hole bands are split away from the heavy bands by strain. Therefore the band edge optical properties will be dominated by heavy hole to conduction band transitions.

## Semiconductor Waveguides-Coupled Modes

For information processing, coupled waveguides are often used. One configuration is the nonlinear directional coupler (NLDC). This was devised for optically activated switching between two output ports. The switching is achieved by using an intensity dependent factor for the refractive index in a linear directional coupler. This locally alters the wave vector mismatch between the guides in such a way that the effective energy transfer and coupling length can be adjusted by varying the input power. Directional coupler switches with optical gain have been investigated in the linear regime.

In the case of intensity dependent all optical switching in semiconductor waveguides, the coupled equations are given below

$$\begin{aligned} &iq_{1z}+\tfrac{1}{2}\beta_2 q_{1tt}+\delta(|q_1|^2+\sigma|q_2|^2)q_1+kq_2 \\ &\qquad =i\gamma_0 q_1-\frac{i}{2}\gamma_2 q_{1tt}-i\beta(|q_1|^2+\sigma|q_2|^2)q_1 \;, \\ &iq_{2z}+\tfrac{1}{2}\beta_2 q_{2tt}+\delta(|q_2|^2+\sigma|q_1|^2)q_2+kq_1 \\ &\qquad =i\gamma_0 q_2-\frac{i}{2}\gamma_2 q_{2tt}-i\beta(|q_2|^2+\sigma|q_1|^2)q_2 \;, \end{aligned} \tag{4}$$

where k represents the linear cross coupling and the term on the right hand side is due to two photon absorption effects.

In Eq. (4) the intensity dependent nonlinearity is a nonresonant nonlinearity. Resonant nonlinearities can be extremely large, but recovery times are slow because thay are mediated by carrier recombination. However, nonresonant nonlinearities can have rapid recovery times because they are not associated with population transfers (19). Measurements of these nonlinearities can be difficult because several processes including virtual and real carriers and thermal effects can contribute. Several experimental techniques can be used to measure the intentisity dependent nonlinear index changes including four-wave mixing (20), nonlinear Fabry-Perot (21), and fringe shift interferometry (22).

The third order susceptibility is complex with the intensity dependent effects determined by the real part of this complex function. The imaginary part is comprised of two photon absorption (TPA) where two photons are absorbed. This effect causes significant problems in semiconductor waveguides because it is essentially a nonlinear loss mechanisms. Its effects on coupled waveguides are considered.

In order to investigate the effects of TPA on semiconductor properties we used the numerical beam propagation method. Figure 1 shows the input power required to achieve the maximum transmission as a function of TPA for various coupling parameters; k=0.25 (solid line); k=0.5 ( dashed line); and k=1.0 (dot-dashed line). The figure shows that for all values of the coupling parameter, the required input power increases as the value of TPA increases. This represents a damaging effect because it indicates the extent of the input power lost to TPA. As a result, greater and greater input power is required to achieve the same output. Since the input power must be held to a minimum, the TPA presents a challenging problem.

Figure 2 shows the maximum transmission at the input channel as a function of TPA for various coupling parameters; k= 0.25 (solid line); k=0.5 (dashed line); and k=1.0 (dot-dashed line). The figure shows that the amount of power transmitted declines rapidly as TPA increases.

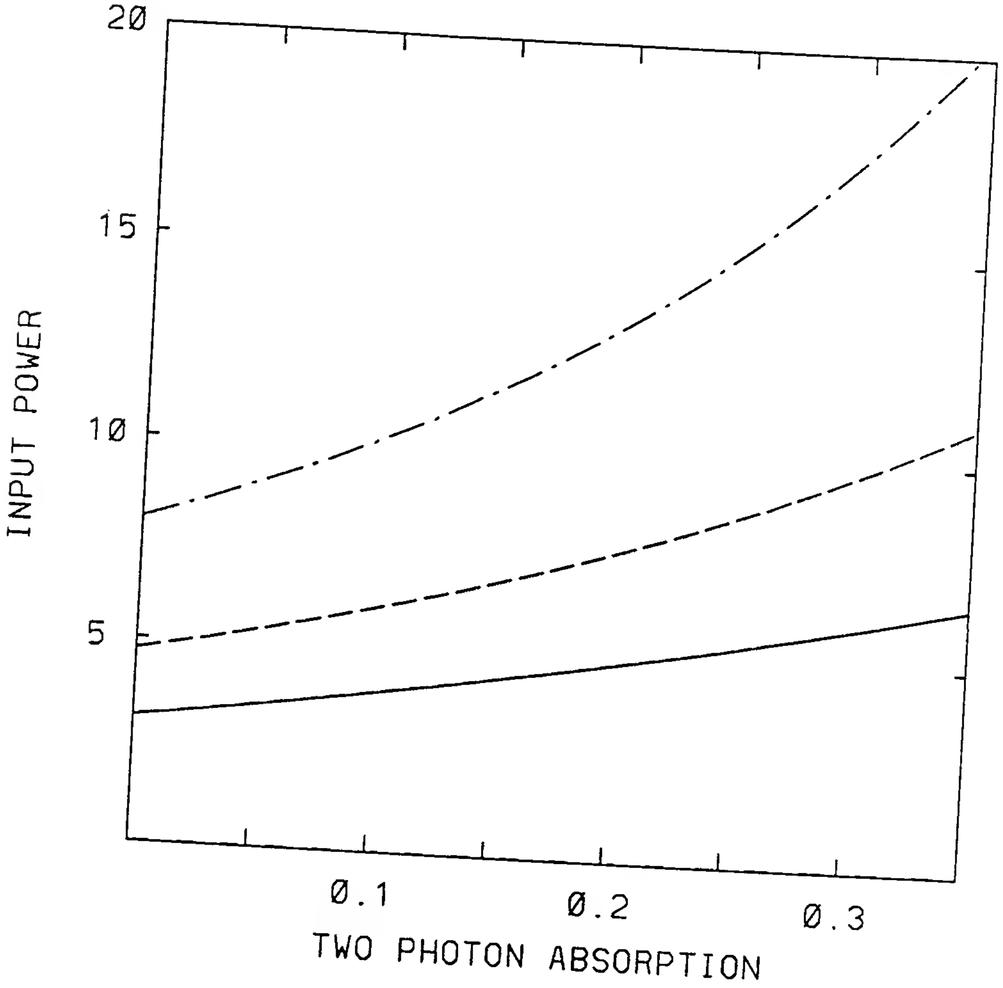


Fig. 1

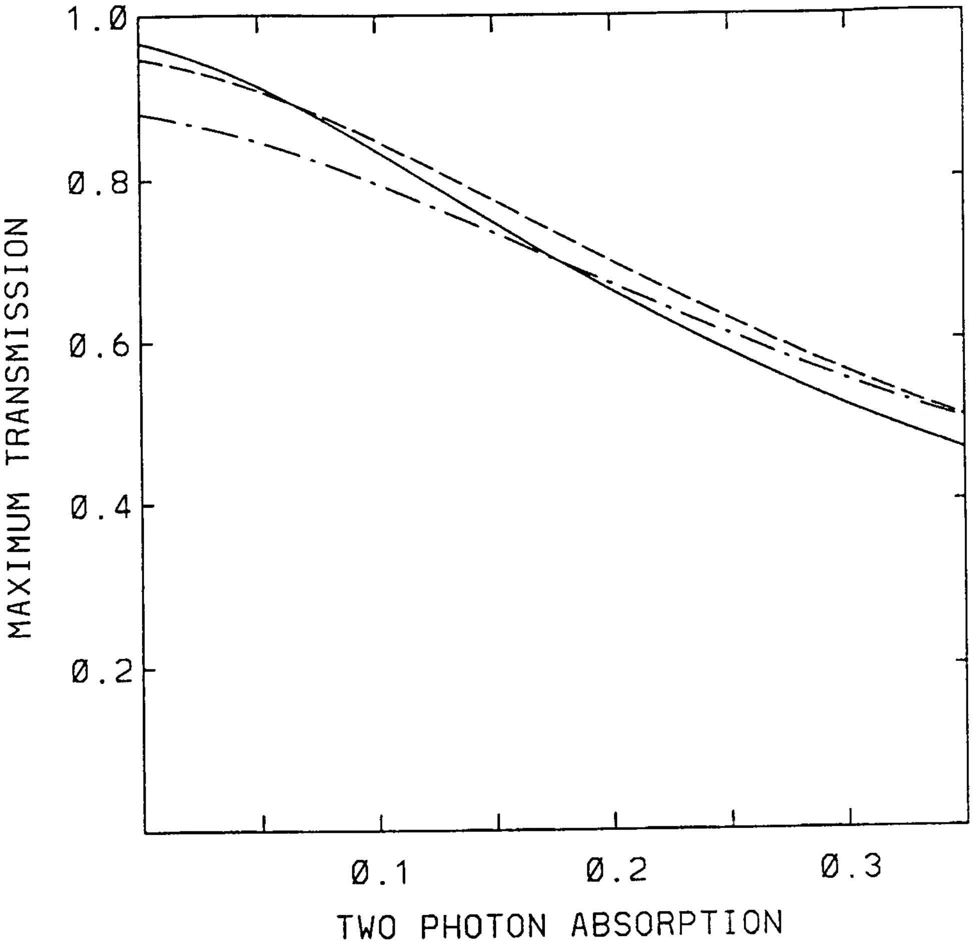

1.Ø
Ø.8
Ø.6
Ø.4
Ø.2
MAXIMUM TRANSMISSION
Ø.1
Ø.2
Ø.3
TWO PHOTON ABSORPTION

Fig. 2

As a result of the TPA and the high powers required, we decided to investigate electro optic effects in semiconductor waveguides. Next a coupled system with distributed gain arising from carrier injection is considered. The active device is composed of two coupled travelling wave semiconductor laser amplifiers. In this system self-switching arises from self-phase-modulation (SPM) associated with optically induced saturation of the gain. A schematic is shown in Fig. 3 below.

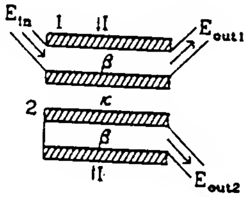


Fig. 3

The effects of TPA and waveguide dispersion are also considered.

Parameters

The subband-to-subband optical gain of the semiconductor is given by (23)

$$g_{nm}(\hbar\omega) = \frac{4\pi^2 e^2 \hbar}{n_0 c m_0^2 W \hbar\omega} \frac{1}{(2\pi)^2} \int d\mathbf{k} \sum_{\sigma} |\, \hat{\epsilon} \cdot \mathbf{P}^{\sigma}_{nm}(\mathbf{k}) \,|^2 \\ \times \delta\left(E^e_n(\mathbf{k}) - E^h_m(\mathbf{k}) - \hbar\omega\right) \left[f^e(E^e_n(\mathbf{k})) - f^h(E^h_m(\mathbf{k}))\right] \qquad (5)$$

where n0 is the index of refraction, c is the speed of light, W is the quantum well width, and the f's are the distribution functions for electrons in the conduction and valence bands. The total material gain is obtained by summing over all subbands n,m and integrating against a Lorentzian function (24)

$$g(E) = \int dE' \sum_{nm} g_{nm}(E')\Delta(E - E'). \tag{6}$$

The radiative current density is evaluated from the dipole transition rate (25) for the quantum confined carrier states

$$\begin{aligned} R_{sp} \;=\; & \int d(\hbar\omega)\frac{4e^2 n_0 \hbar\omega}{3m_0^2 c^3 \hbar^2}\frac{1}{(2\pi)^2}\sum_{nm}\int d\mathbf{k}\sum_{\sigma} \mid \mathbf{P}^{\sigma}_{nm}(\mathbf{k}) \mid^2 \\ & \times\delta\left(E^e_n(\mathbf{k}) - E^h_m(\mathbf{k}) - \hbar\omega\right)\left[f^e(E^e_n(\mathbf{k}))][1 - f^h(E^h_m(\mathbf{k}))\right] .. \end{aligned} \tag{7}$$

At low concentrations the carriers obey the Boltzman statistics with the current being proportional to the square of the injected carrier density n. However at higher carrier densities the emission rate falls off from the n2 curve (23). The light only experiences optical gain in the well region, as a result it is convenient to characterize the gain by a confinement factor given by

$$\Gamma = N\gamma W. \tag{8}$$

where N is the number of quantum wells used in the multiquantum well structure, W is the width of each well and gamma is the optical confinement per unit width of quantum well. A gain curve for InGaAs materials is shown below for 50 A wells for various carrier concentrations (23).

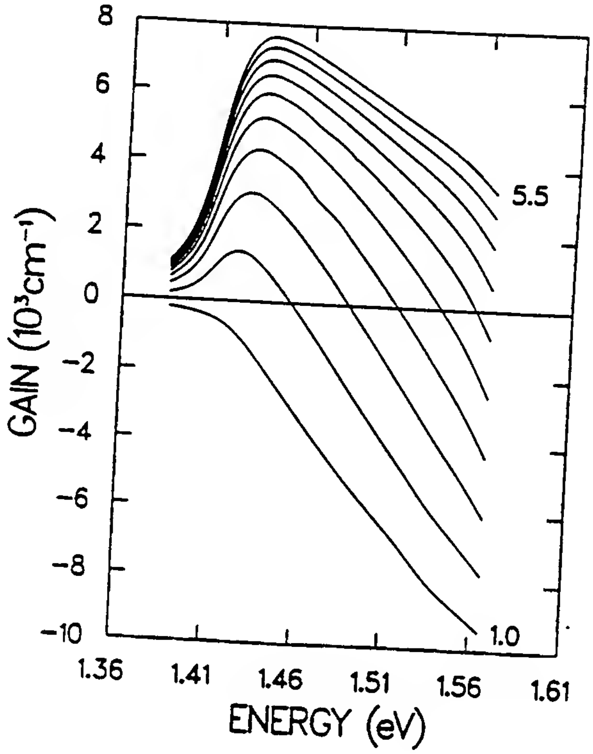


Fig. 4

A Fabry-Perot resonator is assumed for the semiconductor amplifier. The interactions between the carrier concentration and the photon population in the active region of the amplifier can be expressed by a set of coupled multimode rate equations (23)

$$\begin{aligned} \frac{dS_{\hat{e},m}}{dt} &= [N\Gamma g(n,E_m) - \alpha_c](\frac{c}{n_g})S_{\hat{e},m} + \beta N R_{sp}(n) \\ \frac{dn}{dt} &= \frac{J}{eN} - R_{sp}(n) - (\frac{c}{n_g})\sum_{\hat{e},m} \Gamma g(n,E_m)S_{\hat{e},m} \end{aligned} \qquad (9)$$

Mode Equations

An active system is composed of two parallel travelling wave semiconductor amplifiers with distributed gain arising from carrier injection by pumping current into twin stripe contacts. The input beam is injected into channel 1 and the gain of each amplifier is adjusted by varying the pumping current, I. The electromagnetic field is a superposition of the modes of the isolated single-mode waveguides given by

$$E(x,y,z,t) = \frac{1}{2}\sum_{j=1}^{2} [f_j\,(x,y)A_j(z,t) \cdot \exp\{i(\beta z - \omega_0 t)\} + \text{c.c.}] \tag{10}$$

where A is the slowly varying envelope of the electromagnetic field. Linear coupling between the two modes is provided by the evanescent overlap of the two waveguide modes. The coupled equations are

$$\left[\frac{\partial}{\partial z} + \frac{1}{V_g}\frac{\partial}{\partial t}\right]A_1 = i\kappa A_2 + \frac{1}{2}(1 - i\alpha)g_1(N_1)A_1$$

and

$$\left[\frac{\partial}{\partial z} + \frac{1}{V_g}\frac{\partial}{\partial t}\right]A_2 = i\kappa A_1 + \frac{1}{2}(1 - i\alpha)g_2(N_2)A_2. \tag{11}$$

where alpha is the linewidth enhancement factor and the g,s are the intensity gains (dependent on the carrier densities, N). The linewidth enhancement factor is a constant independent of density. When k=0 the expressions reduce to an uncoupled semiconductor amplifier (26-28). For a one coupling length device light will transfer from guide 1 to guide 2 without distortion in the linear limit. When the optical pulse duration exceeds the intraband relaxation time, the induced polarization can be eliminated and the field matter interaction is given by a set of coupled equations. The carrier densities for the semiconductors obeys the equations (27-28)

$$\frac{\partial N_j}{\partial t} = \frac{I_j}{qV} - \frac{N_j}{\tau_c} - \frac{g_j(N_j)}{\hbar \omega_0}|A_j|^2 \tag{12}$$

where I is the injection current, q is the electron charge, and V is the cavity volume. A linear dependence of the gain on the carrier density is given as

$$g_j(N_j) = \Gamma a(N_j - N_0) \qquad j = 1,2 \tag{13}$$

where No is the transparency carrier density.

A cross section averaged gain relaxation equation is used since the waveguide dimensions are smaller than the carrier diffusion length (29)

$$\frac{\partial g_j}{\partial t} = -\frac{(g_j - g_{j0})}{\tau_c} - \frac{g_j}{E_{sat}}|A_j|^2 \qquad j = 1,2 \tag{14}$$

where Esat is the saturation energy, Io is the transparency current and go is the small signal gain. A one coupling length device is considered so that at low input intensities the light exists through the other waveguide. It is assumed that the width of the pulse is much less than the carrier lifetime. As a result the amplifiers are determined by the small signal gain and the linewidth enhancement factor. The experimental value of the linewidth enhancement factor is about 6. If the nonlinear absorption and dispersion are neglected, then the optical switching performance is limited to 50 % of the output energy (30).

Results

In addition to the effects discussed above, dispersive and absorptive factors are important in semiconductor devices. These features are incorporated and their behavior is studied using the numerical beam propagation method (NBP). Figure 5 shows the switching fraction as a function of input intensity. The section on the left hand side corresponds to no dispersion (less than 50% of the light switching); whereas, the right hand side includes the effects of dispersion (about 80% of the light switching). This result represents a significant improvement in waveguide perfromance and is an important consideration for applications and material considerations.

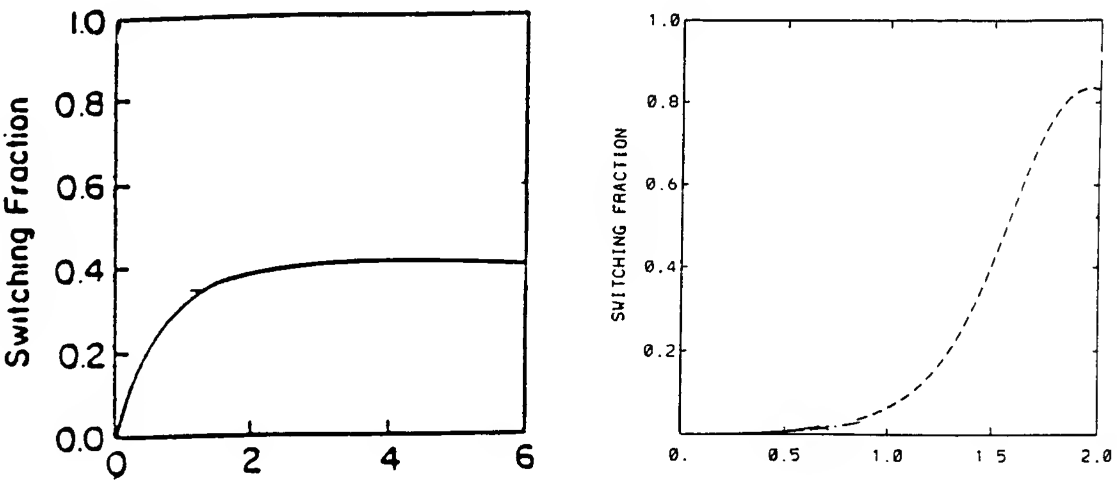


Fig. 5

The effects of TPA can also cause distortion and loss. Fig. 6 shows the effects of high power without the presence of TPA. On the left hand side the light input into one waveguide is amplified and remains largely in that guide. In contrast, the right hand side shows that little light has been transferred. These results show good behavior for the semiconductor materials.

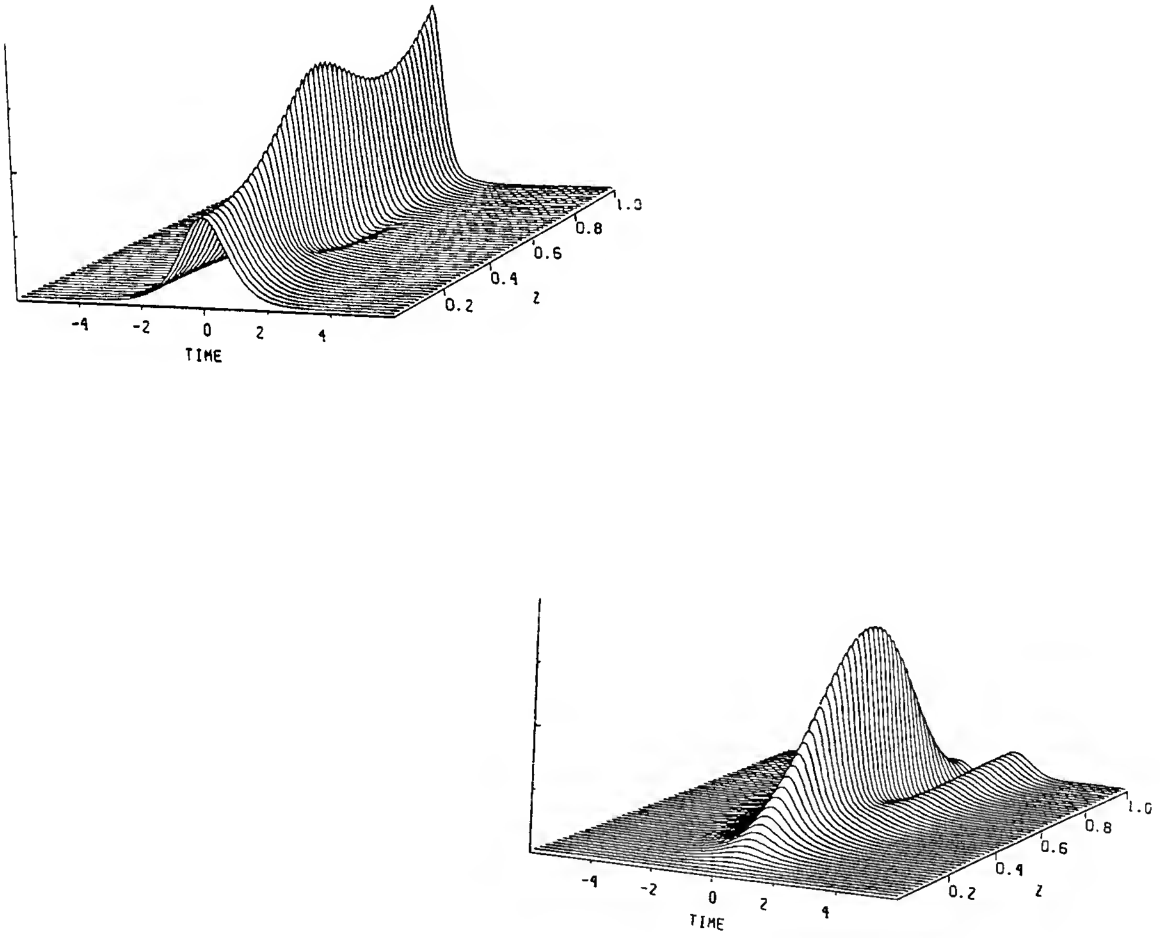


Fig. 6

Next we investigate the effects of TPA on the semiconductors. Figure 7 shows the results when TPA is included. On the left hand side of the figure, the light is amplified and then decays as the absorption takes place. Correspondingly, on the right hand side of this figure more of the light is transmitted than in Fig. 6 and the pulse shape is distorted. However, the amount of distortion and lost power is less than for the configuration using the third order nonlinear optical properties.

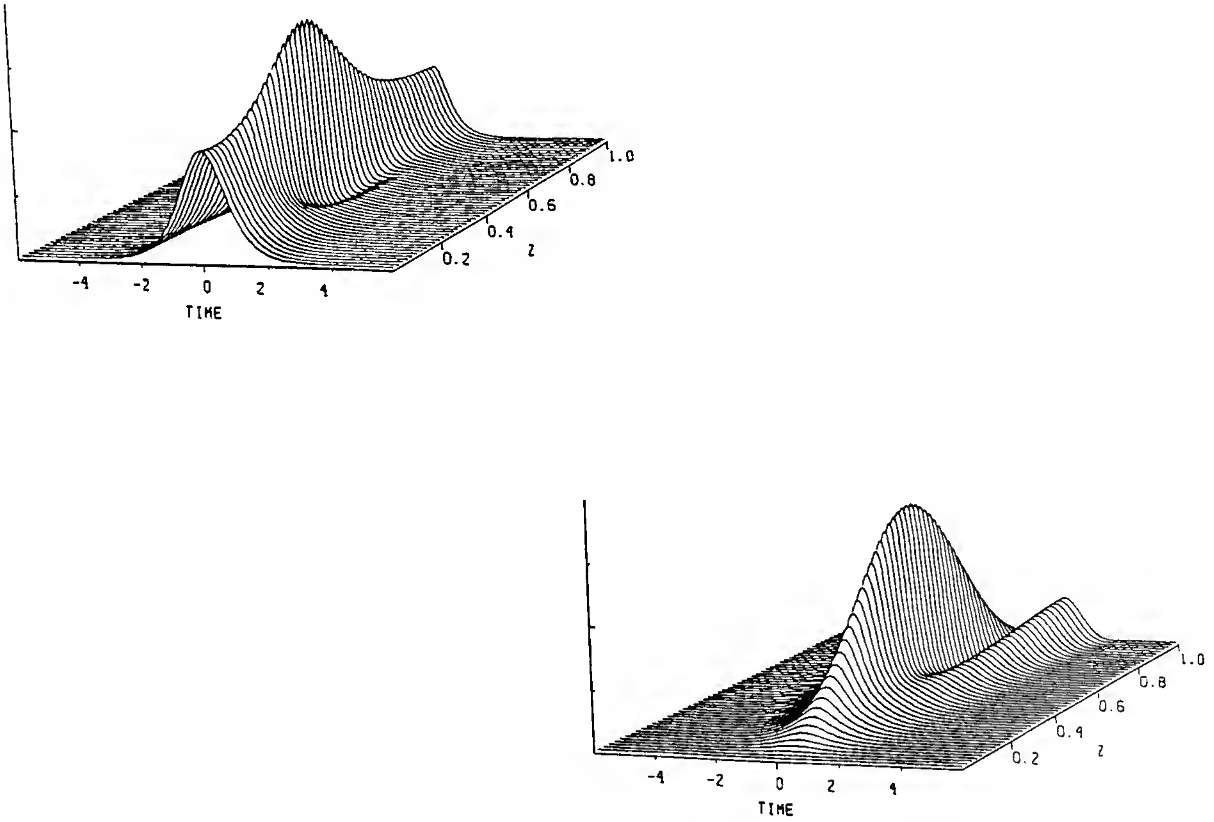


Fig. 7

## Discussion and Summary

We have investigated the properties of third order nonlinear optical and electro-optical systems in semiconductor strained layer materials. These materials are advantageous for many applications in communications and computing. They exhibit band gap engineering in the visible and near infrared spectral regions for use in lasers, modulators, switches and detectors. The materials are also useful for integration of various device components described above. While exhibiting large third order nonlinear optical properties, the effects of two photon absorption degrades the performance of the semiconductor devices. In addition, the third order nonlinear properties require fast laser systems for their investigation. Therefore we also investigated electro-optic effects in semiconductors. In order to avoid the slow time constants of the charge recombination effects due to electron hole recombination rates, we investigated travelling wave amplifiers. These systems have the advantage of relatively low power and perhaps easier integration with other semiconductor devices such as lasers and detectors in a single packaging. It was found that when the dispersive effects of the semiconductors were taken into effect relatively good system performance was achieved. Therefore the travelling wave amplifiers may represent a more advantageous near term application.

Bibliography

1. S. Wabnitz, E.M. Wright, C.T. Seaton, and G.I. Stegeman, Appl. Phys. Lett. 49, 11 (1986).

2. S. M. Jensen, IEEE J. Quant. Electron., QE18, 1580 (1982).

3. Y. Silberberg and G. I. Stegeman, Appl. Phys. Lett. 50, 801 (1987).

4. M. N. Islam, S. P. Dijaili, and J. Gordon, Opt. Lett. 13, 518 (1988).

5. L. Thylen, N. Finlayson, C.T. Seaton, and G.I. Stegeman, Appl. Phys. Lett. 51, 1304 (1987).

6. B. Daino, G.Gregori, and S. Wabnitz, Opt. Lett. 11, 42 (1986).

7. A. Mecozzi, S. Trillo, S. Wabnitz and B. Diano, Opt. Lett. 12, 275 (1987).

8. S. R. Friberg, Y. Silberberg, M.K. Oliver, M.J. Andrejco, M.A.Saifi, and P.W. Smith, Appl. Phys. Lett. 51, 1135 (1987).

9. D.D. Gusovskii, B. Diano, A.A. Maier, V.B. Neustruev, E.I. Shkloskii and I.A. Sheherbakov, Sov. J. Quant. Electron. 15, 1523 (1985).

10. P. LiKam Wa, A. Miller, J.S. Roberts, and P.N. Robson, Integ. Phot. Mtg., 1990, paper WH7.

11. P. Li Kam Wa, P. N. Robson, J. S. Roberts, M.A. Pate, and J. P. David, Appl. Phys. Lett. 52, 2013 (1988).

12. G. Bastard, C. Delalande, Y. Guldner, and P. Voisin, Advances in Electronics and Electron Physics, P.W. Hawkes, ed. (Academic Press, NY, 1988).

13. J. W. Matthews and A.E. Blakeslee, J. Cry. Growth 27, 118 (1974).

14. D.L. Smith and C. Mailhoit, Phys. Rev. Lett. 58, 1264 (1987).

15. C. Mailhoit and D.L. Smith, Phys. Rev. B33, 8360 (1968).

16. S. T. Picraux, L.R. Dawson, G.C. Osbourne an W.K. Chu, Appl. Phys. Lett. 43, 930 (1983).

17. G.C. Osbourne, J. Vacuum Sci. Tech. 21, 459 (1982).

18. K. J. Moore, G. Duggan, K. Woodbridge, and C. Roberts, Phys. Rev. B41, 1090 (1990).

19. M. Yamanishi and M. Kurosaki, IEEE J. Quantum Electron. 24, 325 (1988).

20. W.K. Burns and N. Bloembergen, Phys. Rev. B4, 3437 (1971).

21. Y. H. Lee, A. Chavez-Pirson, S.W. Koch, H. M. Gibbs, S.H. Park, J. Morhange, A. Jeffery, N. Peyghambarian, L. Banyai, A.C. Gossard and W. Wiegmann, Phys. Rev. Lett. 57, 2446 (1986).

22. G. R. Olbright and N. Peyghambarian, Appl. Phys. Lett. 48, 1184 (1986).

23. J. Loehr, Theoretical Studies of Pseudomorphic Quantum Well Optoelectronic Devices, PhD thesis, Univ. of Michigan, Technical Report N. SSEL-202, 1991.

24. G. D. Sanders and Y. C. Chang, Phys. Rev. B35, 1300 (1987).

25. J. J. Sakurai, Advanced Quantum Mechanics (Addison-Wesley, New York, 1967).

26. G.P. Agrawal and N.A. Olsson, Opt. Lett. 14, 500 (1989).

27. G. P. Agrawal and N.A. Olsson, IEEE J. Quantum. Electron. 25, 2297 (1989).

28. N.A. Olsson and G.P. Agrawal, Appl. Phys. Lett. 55, 13 (1989).

29. G. P. Agrawal and N.K. Dutta, Long Wavelength Semiconductor Lasers (Van Nostrand, New York, 1986).

30. S. Trillo, S. Wabnitz, J.M. Soto-Crespo and E.M. Wright, IEEE J. Quantum Electron, 27, 410 (1991).

# DEVELOPMENT OF CONTROL DESIGN METHODOLOGIES FOR FLEXIBLE SYSTEMS WITH MULTIPLE HARD NONLINEARITIES

Final Report
1993 Summer Research Extension Program
RIP # 93-195

Armando Antonio Rodriguez
Assistant Professor
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-7606
(W) (602) 965–3712

Air Force Office of Scientific Research
Bolling AFB
Washington, DC

December 31, 1993

# DEVELOPMENT OF CONTROL DESIGN METHODOLOGIES FOR FLEXIBLE SYSTEMS WITH MULTIPLE HARD NONLINEARITIES

Armando Antonio Rodriguez
Assistant Professor
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-7606
(W) (602) 965-3712

## Abstract

In this research, systematic control design methods have been developed for flexible (distributed parameter) systems with multiple memoryless nonlinearities such as saturating actuators.

## Contents

# DEVELOPMENT OF CONTROL DESIGN METHODOLOGIES FOR FLEXIBLE SYSTEMS WITH MULTIPLE HARD NONLINEARITIES

Armando Antonio Rodriguez
Assistant Professor
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-7606

## 1 Overview and Significance of Research

This report summarizes research conducted under Research Initiation Proposal # 93-195. The research has focussed on two areas: (1) modelling and control of flexible (distributed parameter) systems and (2) control of systems with multiple hard nonlinearities. The major contributions to each area are now summarized.

### 1.1 Modelling and Control of Flexible (Distributed Parameter) Systems

Practically speaking, *flexible systems* are systems for which the structural modes overlap in frequency with the desired bandwidth requirements. Such systems, in general, are modelled by partial differential equations and hence are said to be *distributed parameter* or *infinite-dimensional systems.* Until this work, methods which permit the design of controllers to deliver a pre-specified level of performance for a general distributed parameter (i.e. complex) system did not exist. Throughout the course of this research, methods which permit the design of near-optimal finite-dimensional controllers for such systems have been developed. This has been done for the so-called $\mathcal{H}^\infty$ sensitivity and mixed-sensitivity design paradigms which have received much attention for finite-dimensional systems. The methods have been applied to the problem of controlling a flexible space structure.

### 1.2 Control of Systems with Multiple Hard Nonlinearities

For physical reasons, system designers often want to ensure that certain variables do not exceed pre-specified limits. In the case of missile systems, such variables may include, for example, fin positions, fin rates, angle of attack, sideslip angle, etc. Typically, adhoc modifications are employed and extensive simulations must be performed to justify the modifications. A procedure for systematizing this process has been developed during this research. Unlike previous methods, the method developed here is accompanied by nominal performance guarantees. More specifically, the completed research has shown how an initial control system design can be systematically modified to accommodate memoryless hard nonlinearities (e.g. saturating actuators) which were initially not modelled.

## 1.3 Supporting Publications and Contributions

The following publications acknowledge the support of the Research Initiation Proposal:

### Modelling and Control of Flexible (Distributed Parameter) Systems

1. S.H. Mahloch and A.A. Rodriguez, "System Identification from a Frequency Response," *Proceedings of the 32nd Conference on Decision and Control,* San Antonio, TX, December 15-17, 1993.
2. S.H. Mahloch and A.A. Rodriguez, "System Identification from a Frequency Response: A Sequential Algorithm," submitted for publication in the *Proceedings of the American Control Conference,* Baltimore, MD, June 29–July 1, 1994.
3. A.A. Rodriguez and J.R. Cloutier, "$\mathcal{H}^\infty$ Sensitivity Minimization for Unstable Infinite-Dimensional Plants," *Proceedings of the American Control Conference,* San Francisco, CA, June 2-4, 1993, pp. 2155–2159.
4. A.A. Rodriguez and M.A. Dahleh, "$\mathcal{H}^\infty$ Control of Stable Infinite-Dimensional Systems using Finite-Dimensional Techniques," submitted for publication in *IEEE Transactions on Automatic Control,* 1993.
5. A.A. Rodriguez, "Design of $\mathcal{H}^\infty$ Optimal Finite-Dimensional Controllers for Unstable Infinite-Dimensional Systems," submitted for publication to *AUTOMATICA,* 1993.
6. A.A. Rodriguez and Delano Carter, "Hierarchical HAC$_{\mathcal{H}^\infty}$/LAC Vibration Suppression for a Flexible Space Telescope: SPICE," submitted for publication in the *Proceedings of the American Control Conference,* Baltimore, MD, June 29–July 1, 1994.
7. A.A. Rodriguez and Delano Carter, "$\mathcal{H}^\infty$ Control of SPICE: A Flexible Laser Beam Expander," submitted for publication in the *Journal of Dynamic Systems, Measurements, and Control.*

### Control of Systems with Multiple Hard Nonlinearities

1. A.A. Rodriguez and J.R. Cloutier, "Control of a Bank-to-Turn-Missile with Saturating Actuators," submitted for publication in the *Proceedings of the 1994 American Control Conference,* Baltimore, MD.
2. A.A. Rodriguez and J.R. Cloutier, "Control of a Bank-to-Turn-Missile with Multiple Saturating Actuators," in preparation, to be submitted to *AIAA Journal of Guidance, Control, and Dynamics.*
3. A.A. Rodriguez and S.N. Balakrishnan, "Performance Enhancement for Missile Guidance and Control Systems," proposal submitted for invited session to *1994 American Control Conference,* Baltimore, MD.
4. A.A. Rodriguez and S.N. Balakrishnan, "Performance Enhancement of Missile Guidance Systems in the Presence of Multiple Saturating Actuators," in preparation, Invited session, *1994 AIAA Guidance and Control Conference,* Phoenix, AZ.

5. M. Sonne and A.A. Rodriguez, "PC's in the Design and Evaluation of Guidance and Control Systems for Missiles," to appear in the *Proceedings of the 1994 International Conference on Simulation in Engineering Education*, Tempe, AZ, January 24–26 1994.

6. M. Sonne and A.A. Rodriguez, "A PC-based Graphics System for the Evaluation of Missile Guidance and Control Laws," submitted for publication in the *Proceedings of the American Control Conference*, Baltimore, MD, June 29–July 1, 1994.

7. S.C. Warnick and A.A. Rodriguez, "Longitudinal Control of a Platoon of Vehicles with Saturating Nonlinearities," submitted for publication in the *IEEE Transactions on Control Technology.*

The completed research provides two significant contributions to control system designers. First, it provides a systematic procedure for controlling flexible (distributed parameter) systems. In so doing, it provides a simple method for ascertaining the optimal performance for various $\mathcal{H}^\infty$ criterion. Second, it provides a method for modifying an existing compensator to accommodate initially unmodelled memoryless hard nonlinearities (e.g. saturating actuators, etc.) and maintain, to the extent possible, the directionality properties of the original design.

# 2 Modelling and Control of Flexible (Distributed Parameter) Systems

In this section, the portion of the research results related to modelling and control of flexible (distributed parameter) systems are described.

## 2.1 Modelling: System Identification from a Frequency Response

Often, an analytic model for the system under consideration is not available. Instead, an engineer may have access only to frequency response data. In [1], [2], it is shown how such data can be exploited to develop models which are suitable for control design.

**Iterative Approximation Scheme.** In [1], the authors pose a nonlinear $\mathcal{L}^2$ model-fitting problem which addresses *additive* and *multiplicative modelling errors*. The problem can be stated mathematically as follows:

$$\inf_{d_i, n^{[ij]}} \int_0^{\Omega} tr\{E^H(j\omega)E(j\omega)\} \quad d\omega$$

where,

$$E(s) \overset{\text{def}}{=} \begin{cases} \hat{P}(s) - P(s) & \text{additive;} \\ \hat{P}^{-1}(s)[P(s) - \hat{P}(s)] & \text{multiplicative} \end{cases}$$

$P(jw)$ is a given frequency response (possibly MIMO),

$$\hat{P}(s) \overset{\text{def}}{=} \frac{N(s)}{d(s)}$$

is the approximant to be constructed, and

$$N(s) \overset{\text{def}}{=} \begin{bmatrix} n_{11}(s) & n_{12}(s) & \cdots & n_{1r}(s) \\ n_{21}(s) & n_{22}(s) & \cdots & n_{2r}(s) \\ \vdots & \vdots & \vdots & \vdots \\ n_{r1}(s) & n_{r2}(s) & \cdots & n_{rr}(s) \end{bmatrix}$$

$$n_{ij}(s) \overset{\text{def}}{=} n^{[ij]}_{m_{ij}} s^{m_{ij}} + \cdots + n^{[ij]}_1 s + n^{[ij]}_0$$

$$d(s) \overset{\text{def}}{=} s^n + d_{n-1}s^{n-1} + \cdots + d_0$$

Since the above optimization is nonlinear in the parameters, the problem was reformulated to yield a simpler optimization:

$$\inf_{d_i, n^{[i,j]}} \int_0^{\Omega} tr\{E^H(j\omega)E(j\omega)\} \quad d\omega$$

where now the error $E$ is given by

$$E(s) \overset{\text{def}}{=} W(s)[N(s) - d(s)P(s)]$$

and ideally, to make the simpler problem look like the original problem, one would like to select the weighting function $W$ as follows:

$$W(s) \overset{\text{def}}{=} \begin{cases} \frac{1}{d(s)} & \text{additive;} \\ N^{-1}(s) & \text{multiplicative} \end{cases}$$

However, $d(s)$ and $N(s)$ are unknown *apriori*. Proceeding with algebraic manipulations, the above simpler optimization yields the following quadratic optimization:

$$\inf_x \quad \frac{1}{2}x^T Ax - x^T b + c$$

whose solution can be found by solving a system of linear algebraic equations:

$$Ax = b$$

where $A$, $b$, and $c$ depend on the frequency response data and $x$ contains the unknown parameters of the approximant.

**Iterative Procedure for Weighting.** Since $d(s)$ and $N(s)$ are unknown *apriori*, the following iterative procedure suggests how to select the weighting $W$.

Choose $W_1 = 1$

Solve for $N_1$ and $d_1$

Let $W_2(s) \overset{\text{def}}{=} \begin{cases} \frac{1}{d_1(s)} & \text{additive;} \\ N_1^{-1}(s) & \text{multiplicative} \end{cases}$

```
⋮
Solve for $N_{i-1}$ and $d_{i-1}$
Let $W_i(s) \overset{\text{def}}{=} \begin{cases} \frac{1}{d_{i-1}(s)} & \text{additive;} \\ N_{i-1}^{-1}(s) & \text{multiplicative} \end{cases}$
Solve for $N_i$ and $d_i$
Let $\hat{P}_i = \frac{N_i}{d_i}$
```

Consequently, the initial nonlinear optimization is addressed by solving a sequence of quadratic optimization problems. In [1], the above iterative method is applied to a variety of physical systems, both infinite-dimensional and finite-dimensional. The method is shown to be competitive with existing methods: those which rely on analytic models and those which rely on frequency response data alone.

**Sequential Algorithm for Large Approximants.** A limitation of the previous method is seen when large order approximants are sought. In such a case it requires that one be able to solve a large system of possibly ill-conditioned algebraic equations. Such a method is, of course, limited by the computing resources available. In [2], a sequential method is presented for constructing the approximants - at each step one only needs to solve a small system of equations. The sequential algorithm presented in [2] consequently permits the construction of high-order models which can approximate the original data as closely as desired. The algorithm can be described as follows.

Suppose that an approximant $\hat{P}$ which approximates $P$ over the frequency range $[\Omega_0, \Omega_N] \subseteq R$ is desired. Partition $[\Omega_0, \Omega_N]$ as follows: $[\Omega_0, \Omega_N] = \cup_{i=0}^{N-1}[\Omega_i, \Omega_N]$ where $\Omega_i < \Omega_{i+1}$ for all $i = 0, 1, \ldots, N-1$. To obtain the approximant $\hat{P}$, one proceeds as follows:

Step 1: Obtain approximant of $G(jw)$, denoted $\hat{P}_1(jw)$, on $[\Omega_{N-1}, \Omega_N]$.
Step 2: Obtain approximant of $\Big(P(jw) - \hat{P}_1(jw)\Big)$, denoted $\hat{P}_2(jw)$, on $[\Omega_{N-2}, \Omega_N]$.

$$\vdots$$

Step i: Obtain approximant of $(P(jw) - \sum_{k=1}^{i-1} \hat{P}_i(jw))$, denoted $\hat{P}_i(jw)$, on $[\Omega_{N-i}, \Omega_N]$.

$$\vdots$$

This process is continued up to and including $i = N$. The final approximant is then given by

$$\hat{P}(s) \overset{\text{def}}{=} \sum_{i=1}^{N} \hat{P}_i(s)$$

**Damped Euler-Bernoulli Beam.** The above methods were applied in [2] to a model for a Damped Euler-Bernoulli Beam. The model for a Damped Euler-Bernoulli Beam is given by the following partial differential equation:

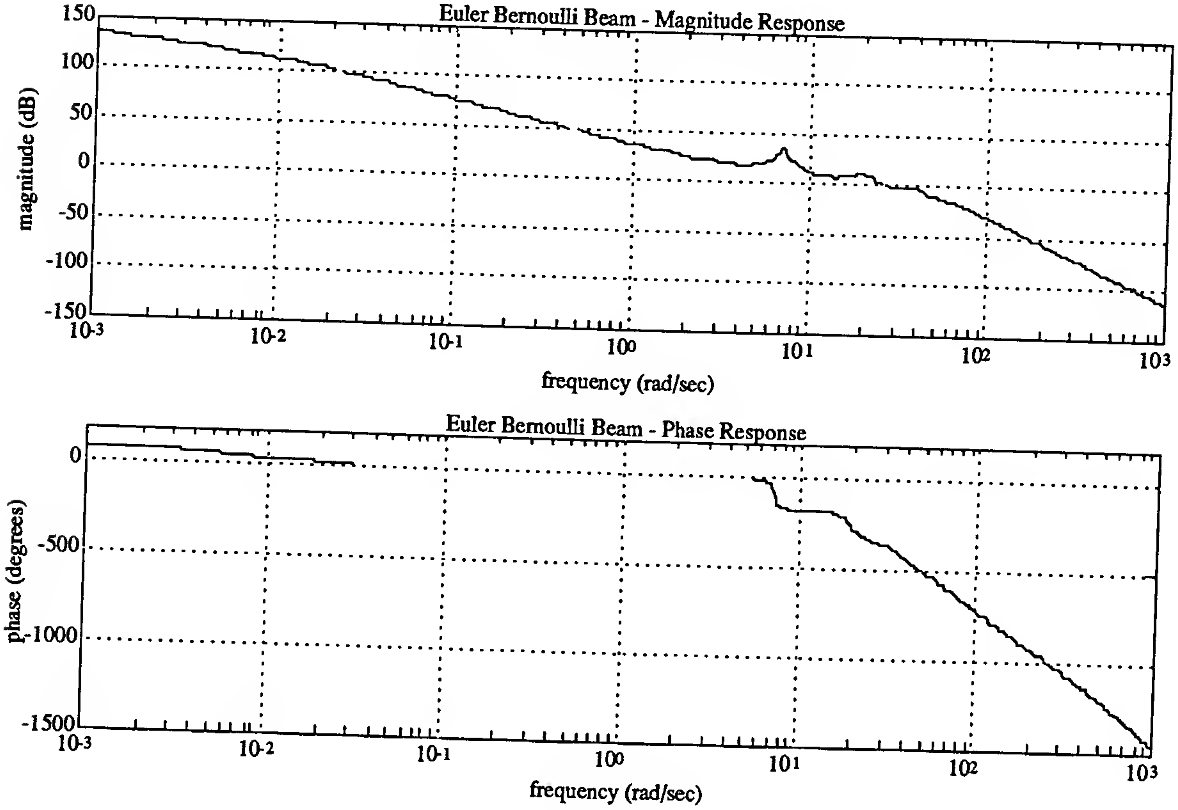


Figure 1: Frequency Response for Damped Euler-Bernoulli Beam

$$\frac{\partial^2}{\partial x^2}\left(E_I\frac{\partial^2 v}{\partial x^2} + c_s I\frac{\partial^3 v}{\partial x^2 \partial t}\right) + m\frac{\partial^2 v}{\partial t^2}\, c\frac{\partial v}{\partial t} = p$$

The associated transfer function is given by

$$P(s) = \frac{v(0,s)}{T(s)} = \frac{cos(a) - cosh(a)}{E_I a^2(cos(a)cosh(a) - 1)}$$

where

$$a^4 = -\frac{ms^2 + cs}{E_I + \frac{c_s I}{s}}$$

The beam parameters used were as follows: $c = c_s = 0.001$, $m = I = 0.1$, $E_I = 0.01$. Figure 1 exhibits the frequency response of the beam.

Figure 2 shows the additive modelling errors which result in generating an [n-1, n] approximant [1] for different values of $n$ when the iterative procedure is used. Figure 3 shows the additive modelling error which results when applying the sequential algorithm to generate a 19th order approximant. Whereas the original iterative method could only (reliably) generate 10th order approximants on a SPARC II workstation, the sequential algorithm easily generated a 19th order approximant. One sees that the peak error for the 8th order model lies below -20 db whereas the peak error for the 19th order model lies below -30 db.

[1] That is, an approximant with an $n^{th}$ order denominator and an $(n-1)^{st}$ order numerator.

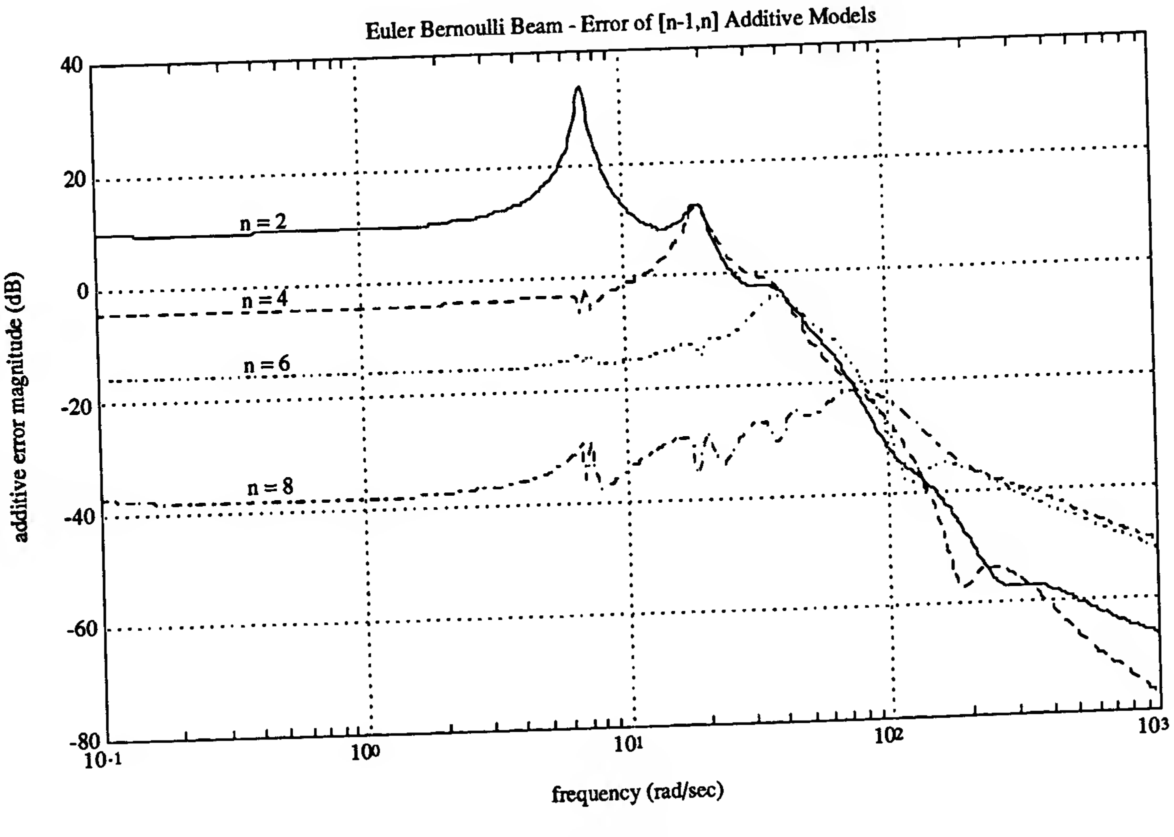


Figure 2: Additive Modelling Error for [n-1, n] Damped Euler-Bernoulli Approximants

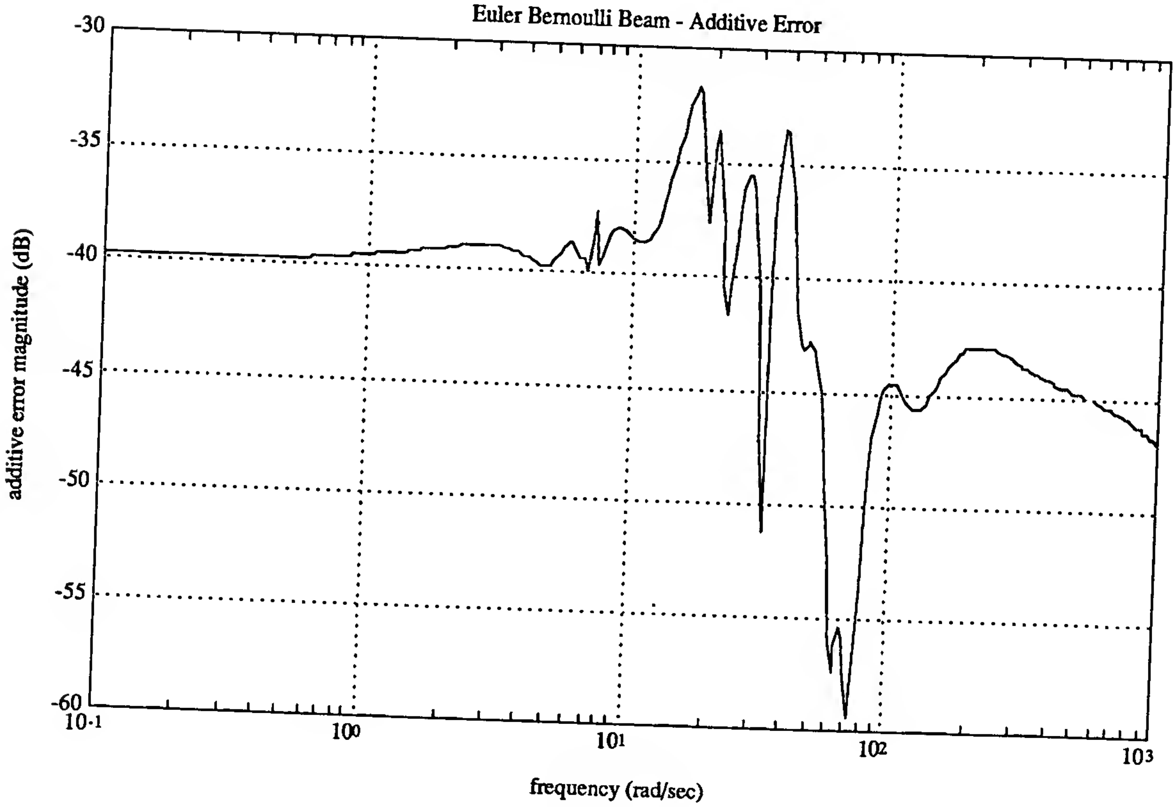


Figure 3: Additive Modelling Error for 19th order [n-1, n] Damped Euler-Bernoulli Approximant

## 2.2 $\mathcal{H}^\infty$ Control for Distributed Parameter Systems

During the 1970's the predominant control design paradigm relied on solving so-called $\mathcal{H}^2$ optimization problems. During the 1980's the focus was on $\mathcal{H}^\infty$ optimization problems [3]. The primary emphasis over the past 20 years has been on methods for finite-dimensional systems. Not until the mid 1980's did infinite-dimensional systems receive appreciable attention by the control systems community (see [4] and references therein). Even then, most researchers focussed on obtaining closed form solutions to complex optimization problems for specific infinite-dimensional systems. Most results obtained were not readily applicable to real systems. Not until [4]-[7], were methods presented which applied to a large class of infinite-dimensional systems. In [4], the authors show explicitly how to construct near-optimal finite-dimensional compensators for a large class of distributed parameter systems subject to $\mathcal{H}^\infty$ design specifications. The main ideas are as follows. For simplicity, it will be assumed that the plant (i.e. system to be controlled) is a linear time invariant (LTI) $\mathcal{L}^2$ finite-gain stable system [8].

**$\mathcal{H}^\infty$ Mixed-Sensitivity Performance Criterion.** Suppose that the *optimal performance* is defined in terms of the following weighted $\mathcal{H}^\infty$ *mixed-sensitivity* problem:

$$\mu_{opt} = \inf_{K \text{stabilizing}} \left\| \frac{\begin{bmatrix} W_1 \\ W_2 K \end{bmatrix}}{1 - PK} \right\|_{\mathcal{H}^\infty}$$

Here, it is assumed that the weighting functions $W_1$ and $W_2$ are stable, minimum phase, proper, and finite-dimensional; i.e. $W_1, W_1{}^{-1}, W_2, W_2{}^{-1} \in \mathcal{RH}^\infty$ [9]. Given this, the above optimization is still infinite-dimensional and hence difficult to solve directly for arbitrary infinite-dimensional plants $P$. For this reason an *Approximate/Design* philosophy is proposed in [4], in which the infinite-dimensional plant $P$ is first approximated by a finite-dimensional approximant $P_n$. Then, one considers the following finite-dimensional optimization

$$\mu_n = \inf_{K \text{stabilizing}} \left\| \frac{\begin{bmatrix} W_1 \\ W_2 K \end{bmatrix}}{1 - P_n K} \right\|_{\mathcal{H}^\infty}$$

for which near-optimal solutions $K_n$ can easily be constructed [3]. In [4], it is shown that if $P_n$ is sufficiently close to $P$, then the *actual performance*, given by

$$\tilde{\mu}_n = \left\| \frac{\begin{bmatrix} W_1 \\ W_2 K_n \end{bmatrix}}{1 - PK_n} \right\|_{\mathcal{H}^\infty}$$

will be close to the optimal performance $\mu_{opt}$. More precisely, if

$$\lim_{n \longrightarrow \infty} \|P_n - P\|_{\mathcal{H}^\infty} = 0$$

then

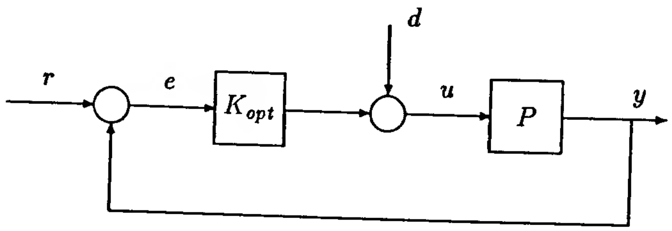


Figure 4: Optimal Infinite-Dimensional Feedback Loop: $\mu_{opt}$

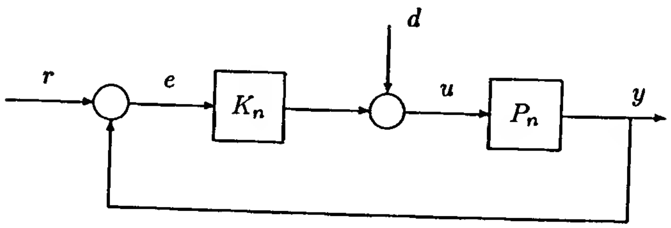


Figure 5: Purely Finite-Dimensional Feedback Loop: $\mu_n$

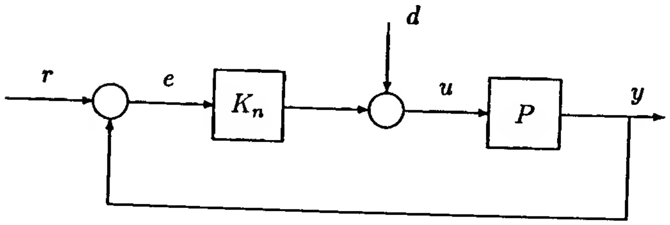


Figure 6: Actual Near-Optimal Feedback Loop: $\tilde{\mu}_n$

$$\lim_{n \longrightarrow \infty} \tilde{\mu}_n = \mu_{opt}$$

Given this, the above *Approximate/Design* procedure provides control engineers with a systematic tool for designing finite-dimensional controllers for complex stable distributed parameter systems (see Figures 4-6). Also, it is interesting to note that the approximants $P_n$ can be constructed directly from frequency response data $P(jw)$. The unstable case is treated in [5], [6].

**$\mathcal{H}^\infty$ Sensitivity Performance Criterion.** Also treated in [4] is the so-called weighted $\mathcal{H}^\infty$ *sensitivity* problem. For this problem, the *optimal performance* is defined as follows:

$$\mu_{opt} = \inf_{K \text{stabilizing}} \left\| \frac{W}{1 - PK} \right\|_{\mathcal{H}^\infty}$$

where $W \in RH^\infty$. For this problem, the construction of a near-optimal compensator is more complex. For this problem, the approximant $P_n$ must be such that the inner and outer parts $P_{n_i}$, $P_{n_o}$ appropriately approximate the inner and outer parts of the plant $P_i$, $P_o$. The unstable case is treated in [5], [6].

# 3 Performance Enhancement for Systems with Multiple Hard Nonlinearities

In this section a method is presented for enhancing the performance of a control system in the presence of multiple memoryless nonlinearities. For simplicity, the discussion is limited to control saturation.

## 3.1 Method for Accommodating Saturating Actuators

While an AFOSR Research Associate at Eglin Air Force Base and throughout the course of this research, the principal investigator has focussed on the problem of enhancing performance in the presence of memoryless hard nonlinearities and, in particular, multiple saturating actuators [10]-[13]. The methods developed are based on the work of [14] and the more recent work of [15]. Other approaches are described in [16], but they suffer from performance and stability problems. To describe the procedure, some notation and assumptions will be needed.

Let $P$ denote a linear time invariant (LTI), multiple-input multiple-output (MIMO) plant. Let $K$ denote a LTI MIMO compensator with state $x(t)$ and state space triple $[A, B, C]$; i.e.

$$\dot{x} = Ax + Be \qquad\qquad u = Cx \tag{1}$$

The pair $(P, K)$ can be visualized as shown in Figure 7. The following assumptions will be made on $P$ and $K$.

**Assumption 3.1 (Assumptions on $P$ and $K$)**
It will be assumed that

(1) $P$ is stable.

(2) $K$ has been designed so that the closed loop system in Figure 7 has desirable properties [2].

(3) $K$ is neutrally stable[3].

(4) The pair $(A, C)$ is observable.

■

The case where $P$ is unstable will be discussed subsequently. The compensator $K$ may be designed using any linear design methodology (e.g. $\mathcal{H}^\infty$, $\mathcal{H}^2$, $\mathcal{L}^1$, LQG/LTR, etc.) [3]. If a complex model is available, the methods described earlier (see [4]-[7]) may prove useful.

It is implicitly assumed that the feedback loop in Figure 7 has "nice" properties. Now let $sat(\cdot)$ denote a general operator modelling a saturation in each control channel. Also, without loss of generality, it will be assumed that each control saturates at ±1 and that each saturation has a transfer characteristic with unity slope [8]. Now consider the feedback loop in Figure 8. It is also implicitly assumed that the performance of this loop is undesirable because of the presence of the saturations. The goal then is to modify $K$ to improve performance. Toward this end, the structure in Figure 9 is proposed. In this figure,

$$u_p(t) = sat(u(t)) \qquad\qquad u(t) = k(t) * [\lambda(x,e)e(t)] \tag{2}$$

where $k(t)$ is the impulse response matrix of the compensator $K$, $*$ denotes convolution, and $\lambda = \lambda(x, e) \in [0,1]$ is a nonlinear scalar gain which depends on the compensator state $x(t)$ and the error signal $e(t) = y(t) - r(t)$. Figure 9 represents a nonlinear system. This is in contrast to Figure 7, where $u_p(t) = u(t)$ and linearity is assured.

In [14], a procedure for computing $\lambda$ is given. The idea behind the procedure is simple. If the system is not saturated, it should be allowed to operate linearly as intended with $\lambda = 1$. If the system is on the "verge of saturation" reduce the gain $\lambda$. Since $\lambda$ is a scalar, such gain reduction preserves the relative coordination of the controls (i.e. the directionality properties of the original design). The procedure requires a state space representation of the compensator and guarantees $\mathcal{L}^\infty$ finite-gain stability [8].

[2] e.g. robust performance, etc.

[3] The matrix A may have eigenvalues on the imaginary axis so long as each has an associated eigenvector; i.e. the geometric and algebraic multiplicities coincide. Repeated roots with geometric deficiency are not permitted here. This assumption is not restrictive for missile autopilots.

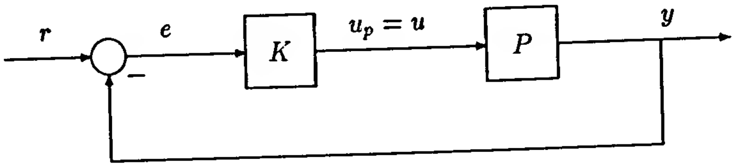


Figure 7: Visualization of Nominal Closed Loop System

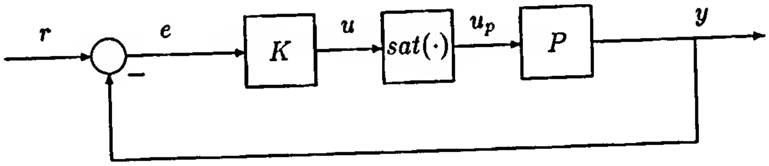


Figure 8: Visualization of Compensator in System with Multiple Saturating Actuators

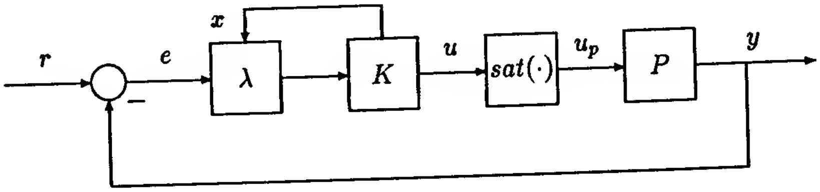


Figure 9: Visualization of Modified Compensator in System with Multiple Saturating Actuators

To present the procedure, the following definition is necessary.

**Definition 3.1** Suppose $A \in \mathcal{R}^{n \times n}$. Given this, define the function $g : R^n \longrightarrow R_+$ as follows:

$$g(x) \overset{\text{def}}{=} \left\| Ce^{At}x \right\|_{\mathcal{L}\infty}$$

■

Notice that this function depends entirely on the homogeneous (unforced) response of the compensator, $K = [A, B, C]$. Given this, it is also useful to define the following set:

**Definition 3.2**

$$B_{\text{AC}} \overset{\text{def}}{=} \{x \in R^n : g(x) \leq 1\}$$

■

With this definition, it follows that if the initial compensator state $x(0)$ lies within or on the boundary of $B_{AC}$ and $\lambda$ is set to zero, then the controls $u = Cx = Ce^{At}x(0)$ cannot grow in magnitude above unity.

The following proposition contains many useful properties of the function $g$ and the set $B_{AC}$.

**Proposition 3.1 (Properties of $g$ and $B_{AC}$)**

(1) $g$ is finite-valued, positive homogeneous, radially non-decreasing.
(2) $g$ is subadditive, convex, continuous, and defines a cone.
(3) $B_{AC}$ is compact and convex.

■

A consequence of $g$ being positive homogeneous is that $g(x) = \|x\| g(\frac{x}{\|x\|})$. This implies that $g$ is completely determined from its values on the unit sphere.

A continuous-time algorithm for constructing $\lambda$ can now be given.

**Algorithm 3.1 ( Construction of $\lambda$ )**
Let $x$ denote the state of the (continuous-time) compensator $K$ at time $t$. Let $e$ denote the error signal at time $t$. The following continuous-time algorithm is proposed for constructing $\lambda$ at each time $t$.

(i) If $x$ lies within $B_{AC}$, then $\lambda = 1$.

(ii) If $x$ lies on the boundary of $B_{AC}$, then maximize $\lambda \in [0, 1]$ such that

$$\lim_{\epsilon \to 0+} sup \frac{g(x + \epsilon[Ax + B\lambda e] - g(x)}{\epsilon} \leq 0 \qquad (3)$$

(iii) If $x$ lies outside $B_{AC}$, choose $\lambda \in [0, 1]$ such that above expression is minimized.

■

It should be noted that the expression in equation (3) is essentially the time derivative of $g$ along the trajectories of the modified compensator:

$$\dot{x} = Ax + B\lambda(x,e)e \qquad\qquad u = Cx \tag{4}$$

given in Figure 9. More precisely, since $g$ in general is not differentiable, the limit in equation (3) denotes the ***upper right Dini derivative*** - a quantity which is well defined for $g$.

To implement the algorithm, one must be able to determine where the compensator state $x$ lies with respect to the boundary of $B_{AC}$. To do this, one must be able to evaluate $g$ on-line. Issues associated with this will be discussed in the following section. However, given that $\lambda$ is computed in accordance with Algorithm 3.1, one obtains the following closed loop performance guarantees.

**Theorem 3.1 (Guaranteed Closed Loop Properties)**

Suppose that $\lambda$ is constructed in accordance with Algorithm 3.1. Let $x_0 = x(0)$. Given this, each of the following holds.

(1) If $x_0$ lies within $B_{AC}$, then $\|u(t)\|_{\mathcal{L}^\infty} \leq 1$ for all $e$.

(2) If $x_0$ does not lie within $B_{AC}$, then $\|u(t)\|_{\mathcal{L}^\infty} \leq g(x_0)$ for all $e$.

(3) The closed loop system in Figure 9 will be $\mathcal{L}^\infty$ finite-gain stable.

■

It should be noted that finite-gain stability can be proved because for sufficiently small exogenous signals the system in Figure 9 exhibits linear behavior.

## 3.2 Computational Issues

As pointed out in the previous section, the function $g$ must be evaluated on-line. This issue is complicated by the fact that the definition for $g$ given in definition 3.1 is not immediately suitable for on-line computations. What is needed is a useful characterization, or approximation, for the function $g$. Also, because Algorithm 3.1 will ultimately be implemented on a digital computer, a discretized version of the algorithm is needed. These points are now addressed.

Suppose that $K$ has a discrete-time realization $[\tilde{A}, \tilde{B}, \tilde{C}]$ as follows:

$$x_{n+1} = \tilde{A}x_n + \tilde{B}e_n \qquad\qquad u_n = \tilde{C}x_n \tag{5}$$

where $\tilde{A}$ is also at least neutrally stable. Given this, $g$ may be approximated as follows.

**Proposition 3.2 (Approximation for $g$.)**
Given that $\tilde{A}$ is at least neutrally stable, it follows that $g$ may be approximated as follows:

$$g(x) \approx \tilde{g}(x) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \tilde{C}\tilde{A}^2 \\ \tilde{C}\tilde{A}^3 \\ \vdots \\ \tilde{C}\tilde{A}^k \end{bmatrix} x \right\|_{\mathcal{L}\infty}$$

where $k$ is some sufficiently large integer which can be determined off-line.

■

It should be pointed out that some of the terms in the above approximation may be unnecessary. To identify these terms, the linear-programming ideas in [15] may prove helpful. Given this, one can implement Algorithm 3.1 as follows.

**Algorithm 3.2 ( Discrete Performance Enhancement Algorithm )**
Let $x_n$ denote the state of the (discrete-time) compensator $K$ at time $n$. Let $e_n$ denote the error signal at time $n$. The following discrete-time algorithm is proposed for constructing $\lambda_n$ at each $n$.

(i) If $\tilde{g}(x_n) < 1$, then $\lambda_n = 1$.

(ii) If $\tilde{g}(x_n) = 1$, then maximize $\lambda_n \in [0, 1]$ such that

$$\tilde{g}(\tilde{A}x_n + \tilde{B}\lambda_n e_n) - \tilde{g}(x_n) \leq 0 \tag{6}$$

(iii) If $\tilde{g}(x_n) > 1$, choose $\lambda_n \in [0, 1]$ such that above expression is minimized.

■

It should be noted that this algorithm requires that an on-line optimization be performed at each time-step. Consequently, efficient optimization routines must be sought. Because $\lambda$ controls the "amount" of error entering the compensator, it is referred to as an *error governor*.

## 3.3 Unstable Operating Points and Other Hard Nonlinearities

Thus far, the focus has been on systems with dynamics that are locally stable. The scheme presented, however, cannot directly be used when the dynamics are locally unstable. Fundamentally, this is because an unstable plant has a finite downward gain margin. To address unstable operating points, the ideas in [14] were considered. In [14], the author introduces a *reference governor* which limits the rate of growth of the reference command so that saturation is avoided. It processes the reference command, the compensator state, and the plant state. Access to all of the plant states, however, is not a practical assumption. Consequently, while at Eglin Air Force Base, the principal investigator worked on the problem of designing an estimator which could generate appropriate estimates of the plant states to be used by the *reference governor*. The ideas in

[17] and [18] were very helpful because they provided insight into the design of a suitable $\mathcal{L}^2$ estimator. Of importance here is realizing the *peaking phenomenon* which occurs in the design of state estimators. From estimation theory, one expects that as the sensor noise decreases, the estimator bandwidth and response improves. However, it may be that certain state estimation errors may grow significantly before they decay rapidly to zero. This peaking phenomenon is discussed in [17] for Linear Quadratic Regulators. From [17], it follows that conditions can be obtained to prevent such peaking from occurring. The condition, loosely speaking, requires that the transfer function matrix from the process noise to the measurements possesses the maximum number of transmission zeros. This will be guaranteed if this transfer function matrix has (normal) rank equal to the rank of its associated first Markov parameter.

It should be noted that an $\mathcal{L}^\infty$ estimator, i.e. one which is accompanied with bounds on the peak estimation errors, is perhaps more appropriate to address the technical issues discussed above. This work is still in progress. Other hard nonlinearities, such as rate limiters and $\alpha$-limiters, can easily be addressed using the estimation methods ideas proposed here.

### 3.4 Extension to Nonlinear Compensators

It should also be noted that the above method extends directly to bilinear compensators [19]; i.e. nonlinear compensators with state space descriptions as follows:

$$\dot{x} = Ax + B(x)e \qquad\qquad u = Cx \tag{7}$$

where $A$ and $C$ are constant matrices and $B(\cdot)$ is a matrix function of the compensator state $x$. This is particularly interesting because nonlinear systems with fading memory can be approximated by bilinear systems.

More generally, the ideas also seem to directly extend to compensators having the form

$$\dot{x} = Ax + B(x,e) \qquad\qquad u = C(x) \tag{8}$$

where $B(x,0) = 0$ and $C(\cdot)$ is "sufficiently well-behaved". These ideas are currently being examined by the principal investigator.

## 4 Applications

### 4.1 Flexible Space Structure: SPICE

In [20], [21], the ideas discussed in section 2 were applied to the problem of controlling a flexible structure which may operate as a space telescope or as a laser beam expander. As a flexible laser beam expander, the structure chemically generates a narrow laser beam and expands it via Cassegrain primary/secondary mirror system. Generation of the high intensity beam and coolant flow through the mirrors result in excitation of the structural modes. The problem here is to perform rapid/accurate slewing/pointing maneuvers without excitation of the flexible modes. Having been developed by the Air Force, with Lockheed and Honeywell as subcontractors, this system physically resides at the Phillips Laboratory on Kirtland Air Force Base. This system has been assigned the acronym SPICE for Space Integrated Control Experiment.

A 260 state NASTRAN [4] model with 18 actuators and 18 sensors was used to model the loaded structure. A 130th reduced order model was used to design a hierarchical control system. Low authority feedback was used to obtain a design plant with increased structural damping. The $\mathcal{H}^\infty$ design methodology was then used to obtain a High authority control law. The resulting design achieves the desired 40 db line-of-sight (LOS) attenuation Air Force specification. Slewing maneuvers and other issues are currently under investigation.

## 4.2 EMRAAT BTT Missile with Saturating Actuators

While an AFOSR Research Associate, and through the support of an AFOSR Research Initiation Award, the principal investigator has studied the problem of enhancing performance for highly maneuverable Bank-to-Turn (BTT) missile systems with multiple saturating actuators [10], [12], [13]. Because of the significance of the results obtained thus far, the principal investigator has organized a special session at the 1994 American Control Conference. The session is entitled *Performance Enhancement for Missile Guidance and Control Systems.* A description of the session has been included in Appendix IX. Some of the key results are now discussed.

**Missile Model.** BTT missiles offer higher maneuverability over conventional Skid-to-Turn (STT) missiles by the use of an asymmetrical shape and/or the addition of a wing [22], [23]. The model used in this study has been for an EMRAAT (Extended Medium Range Air-to-Air Technology) BTT missile. For simplicity, focus has been placed on the yaw/roll dynamics at an operating point with a Mach number of 2.5, a dynamic pressure $Q_{pres} = 1720\ lb/ft^2$, and an angle of attack $\alpha = 20\ degrees$. The model is given by the following system of ordinary differential equations [22]:

$$\dot{x}_p = A_p x_p + B_p u_p \qquad\qquad y = C_p x_p \tag{9}$$

where

$$A_p = \begin{bmatrix} -0.818 & -0.999 & 0.349 \\ 80.29 & -0.579 & 0.009 \\ -2734 & 0.5621 & -2.10 \end{bmatrix} \quad B_p = \begin{bmatrix} 0.147 & 0.012 \\ -194.4 & 37.61 \\ -2176 & -1093 \end{bmatrix} \quad C_p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{10}$$

$$u_p = \begin{bmatrix} \text{rudder} & \text{aileron} \end{bmatrix}^T \quad x_p = \begin{bmatrix} \text{sideslip} & \text{yawrate} & \text{rollrate} \end{bmatrix}^T \quad y = \begin{bmatrix} \text{sideslip} & \text{yawrate} \end{bmatrix}^T \tag{11}$$

and variables are measured in degrees or degrees/second. This system has poles at $s = -0.6579, -1.4195 \pm j32.1542$ and thus the missile is assumed to be operating near a stable equilibrium. Unstable operating points will be discussed below. For completeness, the singular values of the above missile (plant) transfer function matrix $P(s) = C_p(sI - A_p)^{-1}B_p$ are shown in Figure 10.

[4] NASTRAN is a NASA structural analysis program.

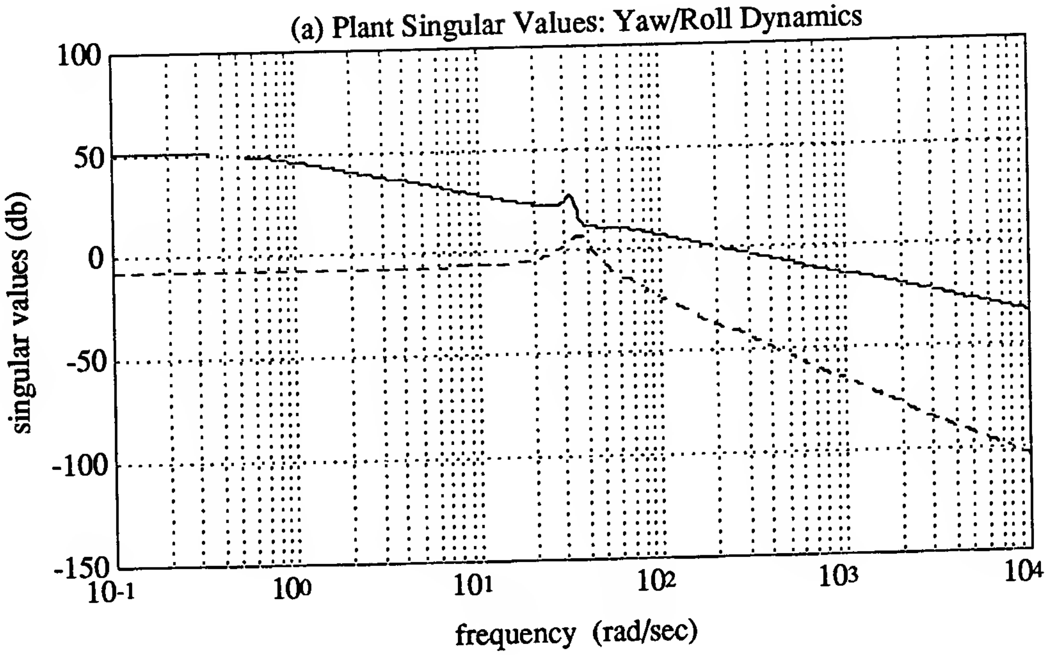


Figure 10: Plant Singular Values

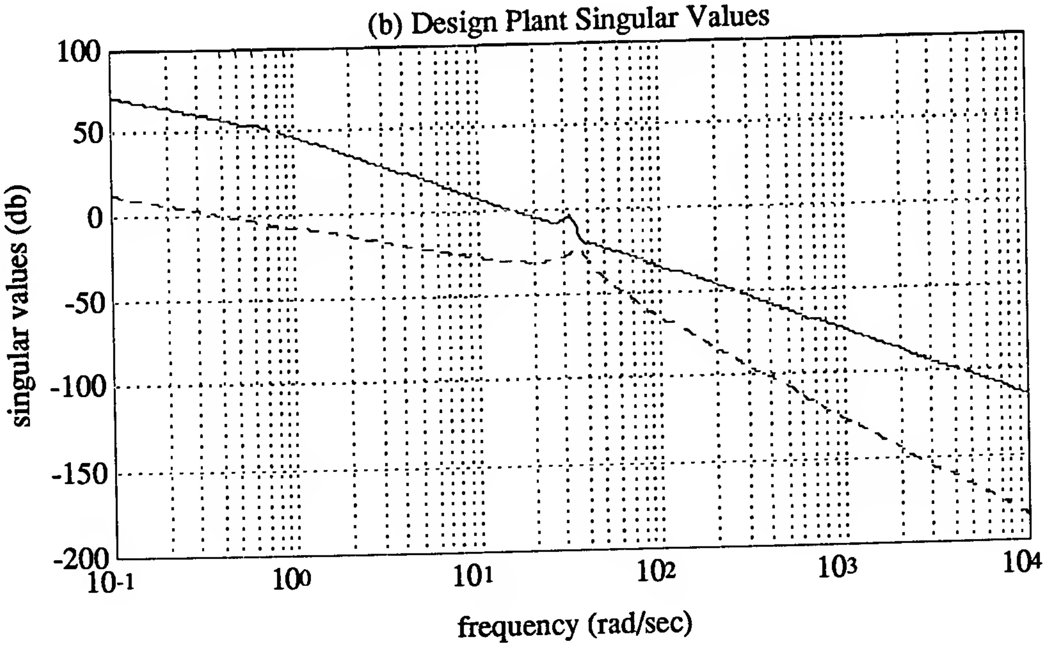


Figure 11: Design Plant Singular Values

**Nominal Autopilot Design.** For purposes of demonstrating the performance enhancement concept, the LQG/LTR design methodology [24] was used to obtain a nominal linear autopilot design. The procedure for obtaining the nominal autopilot design is now described.

**Step 1: Form Design Plant.** To guarantee zero steady state error to step commands, the plant $P = [A_p, B_p, C_p]$, given above, was augmented with integrators; one in each control channel. The resulting system is called the *design plant* and has a state space triple $[A_{des}, B_{des}, C_{des}]$ given by

$$A_{des} = \begin{bmatrix} 0 & 0 \\ B_p & A_p \end{bmatrix} \qquad B_{des} = \begin{bmatrix} I \\ 0 \end{bmatrix} \qquad C_{des} = \begin{bmatrix} 0 & C_p \end{bmatrix} \tag{12}$$

The design plant singular values have been plotted in Figure 11.

**Step 2: Design Target Loop.** The next step in the process is to design the target (desired) open loop transfer function matrix. The target loop was selected to have a state space triple $[A_{des}, H, C_{des}]$ where the *filter gain matrix* $H$ was selected to be

$$H = 10 \;\; [v_4 \quad -v_5] \begin{bmatrix} -0.0003 & 0.0912 \\ -1.6482 & -1.7159 \end{bmatrix}^{-1} = \begin{bmatrix} -0.1036 & -0.0055 \\ -4.8067 & 0.0009 \\ 2.0003 & 0.0000 \\ 0.0003 & 2.0000 \\ 4.8980 & 5.7272 \end{bmatrix} \tag{13}$$

where $v_4 = \begin{bmatrix} 0.0009 & 0 & -0.0001 & -0.3296 & -0.9441 \end{bmatrix}^T$ and $v_5 = \begin{bmatrix} 0 & 0.0440 & -0.0182 & 0.3432 & 0.9381 \end{bmatrix}^T$ are right eignenvectors of $A_{des}$ corresponding to the eigenvalue $\lambda = 0$. This makes the pair $(A_{des}, H)$ uncontrollable because the left eigenvectors of $A_{des}$ associated with the missile modes lie in the left null space of $H$ [25]. By so doing, one obtains a target loop which looks like an "integrator" with gain crossover frequency near 2 rad/sec. For convenience, the target loop singular values have been plotted in Figure 12.

**Step 3: Recover Target Loop.** The next step in the process is to recover the target loop by solving an appropriately formulated *cheap control* problem. This amounts to solving the Control Algebraic Riccati Equation (CARE)

$$0 = K_c A_{des} + A_{des}^T K_c + C_{des}^T C_{des} - K_c B_{des} \frac{1}{\rho} B_{des}^T K_c \tag{14}$$

for the unique symmetric positive definite solution $K_c$. This was done with a *recovery parameter* $\rho = 10^{-6}$. Doing this yields the *control gain matrix*

$$G_\rho = \frac{1}{\rho} B_{des}^T K_c = \begin{bmatrix} 612.4440 & -88.1472 & -410.4327 & -965.3118 & -1.7612 \\ -88.1472 & 164.2489 & -564.4860 & 254.7471 & -7.1359 \end{bmatrix} \tag{15}$$

The final (i.e. nominal) compensator, $K$, is then given by

$$\dot{x} = Ax + Be \qquad u = Cx \tag{16}$$

where $e = r - y$ and

$$A = A_{des} - B_{des} G_\rho - H C_{des} \qquad B = H \qquad C = G_\rho \tag{17}$$

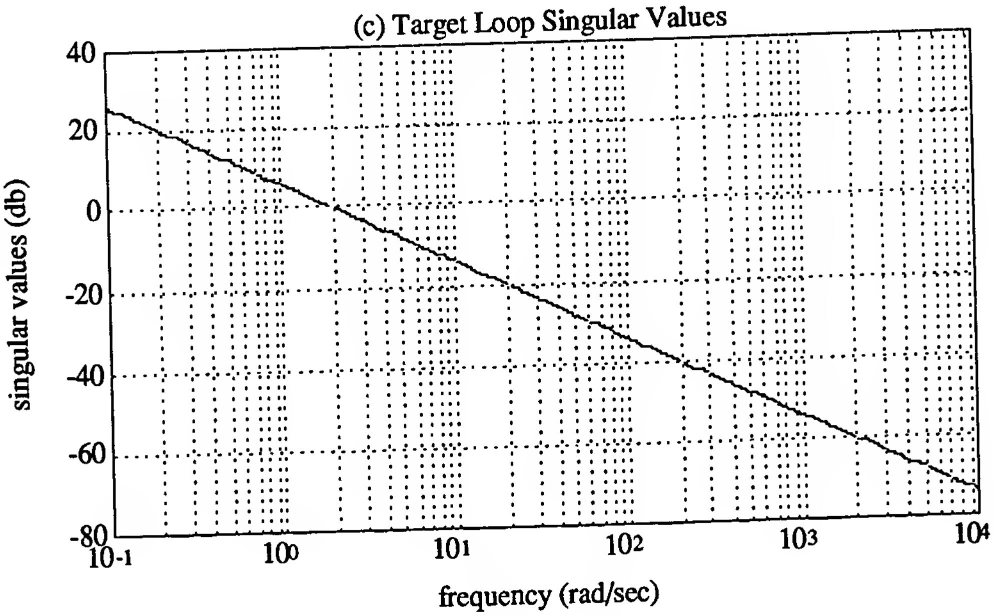


Figure 12: Target Loop Singular Values

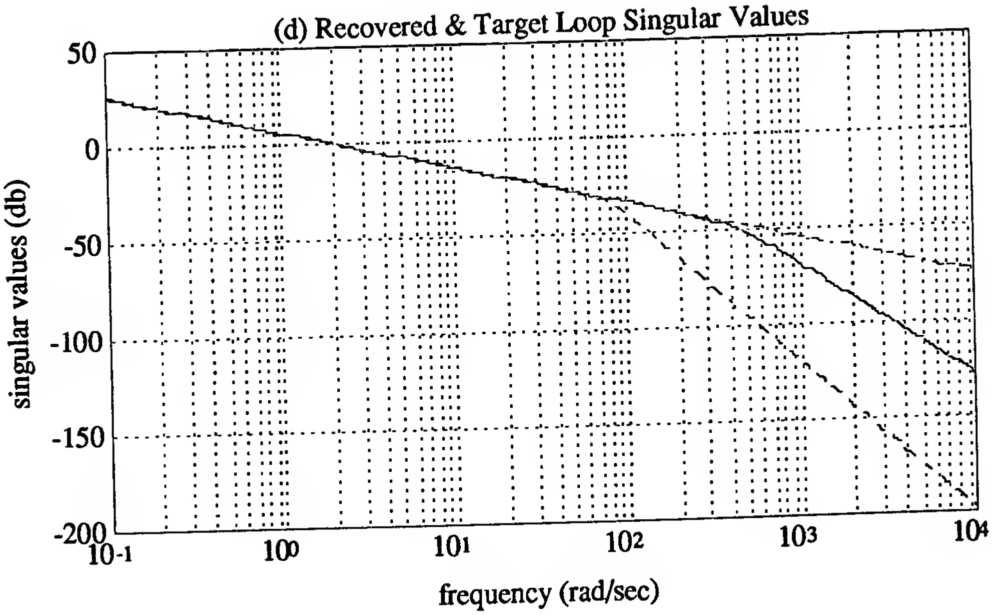


Figure 13: Recovered and Target Loop Singular Values

A balanced realization for $K = [A, B, C]$ is then given by

$$A = \begin{bmatrix} -0.2903 & -107.7837 & 6.6692 & -2.5820 & -0.4031 \\ 107.6795 & -97.8098 & 63.9509 & -4.5172 & -5.3516 \\ -6.7247 & 64.8158 & -54.1931 & -40.7948 & 5.1149 \\ 3.2148 & 2.0979 & 29.5571 & -631.1537 & 429.8894 \\ 0.3648 & -3.3887 & 3.0907 & -460.0290 & -0.7433 \end{bmatrix} \tag{18}$$

$$B = \begin{bmatrix} 2.2840 & 0.4772 \\ -40.7546 & 2.1291 \\ 18.4665 & -0.2215 \\ -2.0715 & -44.6786 \\ -0.9753 & -1.1753 \end{bmatrix} \qquad C = \begin{bmatrix} 0.8562 & 8.5376 & -1.7085 & 43.9089 & 1.1235 \\ 2.1706 & 39.9071 & -18.3886 & -8.5133 & 1.0346 \end{bmatrix} \tag{19}$$

Again, for convenience, the recovered singular values have been plotted in Figure 13.

The performance enhancement scheme discussed in section 3 was applied to the EMRAAT model and LQG/LTR compensator discussed above. The discrete-time realization

$$x_{n+1} = \tilde{A}x_n + \tilde{B}e_n \qquad u_n = \tilde{C}x_n \tag{20}$$

where

$$\tilde{A} = I + T_s A \qquad \tilde{B} = T_s B \qquad \tilde{C} = C \qquad T_s = 0.01 sec \tag{21}$$

was used for $K$. A constant reference command of $r = [4.2 \ \ -4.2]^T$ was selected to evaluate performance with respect to command following. Figure 14 contains the linear responses (see Figure 7) and the responses which occur when saturations are inserted in each control channel (see Figure 8). The saturation limits used were ±8.

As expected, the linear responses are very good. The transient is well behaved and the steady state tracking error is zero. The latter follows from the *Internal Model Principle* and the fact that the compensator has an integrator in each control channel. When the saturations are introduced, however, the integrators in the compensator wind-up. This is seen in the observed aileron response generated by the compensator. The rudder response follows the linear response closely. The sideslip response is not able to achieve the commanded steady state sideslip but it also remains close. More dramatic is the observed yaw rate response which is unable to come close to the commanded yaw rate. It is apparent that the saturations, particularly the one in the aileron control channel, destroys the directionality properties of the original LQG/LTR-based autopilot $K$.

To maintain the original autopilot directionality properties and prevent wind-up, the performance enhancement scheme described above was used (see Figure 9). The resulting regulated responses are given in Figure 15. The unregulated responses are repeated in the figure for comparison sake. It is seen that the scheme maintains the directionality properties of the original autopilot to the extent possible. It permits the system to operate on the edge of saturation and completely eliminates the wind-up effects. The resulting aileron control produced by the modified compensator, for example, reaches the -8 rail and remains there.

### 4.2.1 Graphical Tool for Evaluation of Missile-Target Intercept

In [26], [27] the authors describe a $C_{++}$/Windows based 6 dof developed for a BTT missile to graphically visualize and evaluate missile-target intercepts. The program allows the user to specify different guidance

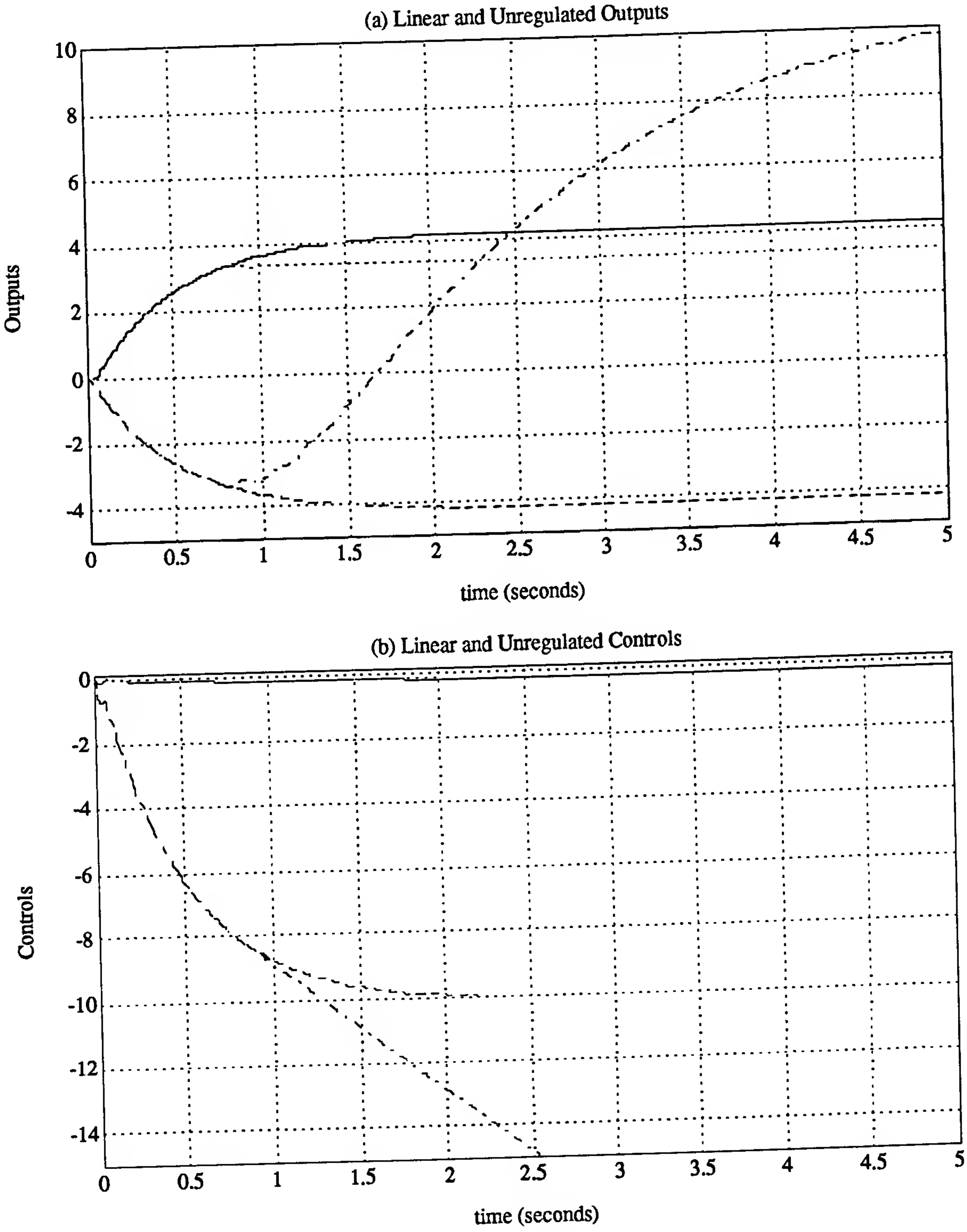


Figure 14: Linear and Unregulated Responses

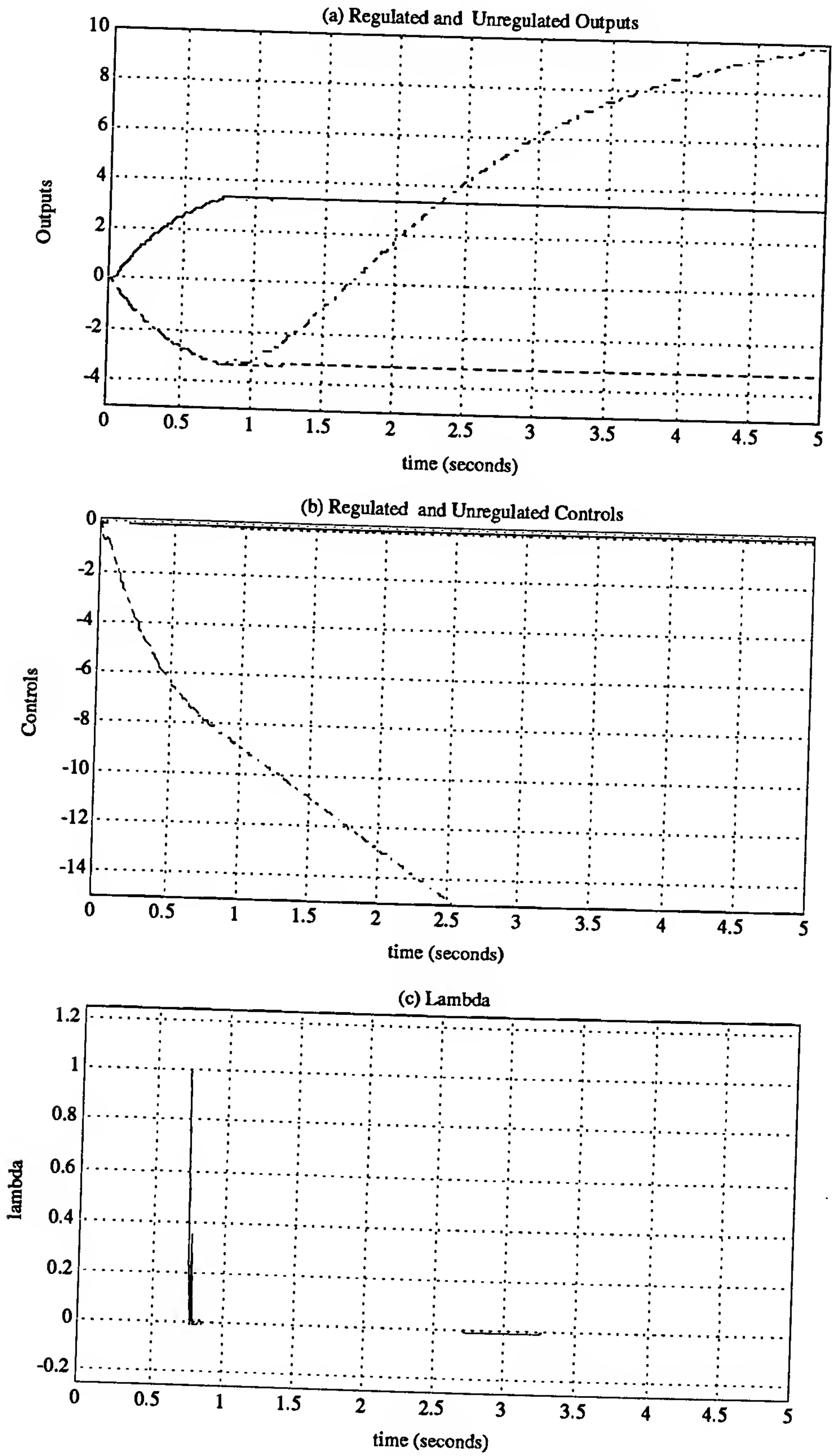


Figure 15: Regulated and Unregulated Responses

laws, autopilots, engagement geometries, and target maneuvers. The program also permits a visual evaluation of the effects of hard nonlinearities on the guidance and control systems. The program is currently being used to visualize the performance of the saturation algorithm described earlier.

### 4.3 Platoon of Vehicles with Saturating Actuators

In [28], the saturation method described in section 3 was applied to the problem of controlling the longitudinal dynamics of a platoon of vehicles with saturating actuators. Each vehicle in the platoon is modelled by a $2^{nd}$ order nonlinear differential equation. One degree of freedom is used to capture the vehicle dynamics; another to capture the engine dynamics. The nominal control law is designed using feedback linearization techniques. The throttle on each vehicle is assumed to be limited and various versions of the algorithm discussed earlier is applied.

### 4.4 Invited Sessions: Missile Guidance and Control

The completed research on memoryless hard nonlinearities and, in particular, on saturating nonlinearities [10], [11] has led to the organization of two invited sessions: one at the *1994 American Control Conference* to be held in Baltimore, MD and one at the *1994 Guidance, Navigation, and Control Conference* to be held in Phoenix, AZ [12], [13]. Both address performance enhancement and integrated design for missile guidance and control systems.

## 5 Summary and Directions for Future Research

In summary, two significant contributions have been made in this research. The first is a systematic design methodology for general distributed parameter systems. The second, is a procedure which permits control engineers to directly take into account memoryless hard nonlinearities such as saturating actuators, rate limiters, etc. The procedure enhances performance in the presence of such nonlinearities, systematizes the design process, and is computationally feasible with the computing resources available on existing systems.

The research is continuing as follows. New performance criterion (other than $\mathcal{H}^\infty$) are being considered for distributed parameter systems. Also, the performance enhancement scheme is being extended to more general nonlinear compensators.

## 6 Bibliography

### References

[1] S.H. Mahloch and A.A. Rodriguez, "System Identification from a Frequency Response," *Proceedings of the 32nd Conference on Decision and Control*, San Antonio, TX, December 15-17, 1993.

[2] S.H. Mahloch and A.A. Rodriguez, "System Identification from a Frequency Response: A Sequential Algorithm,"submitted for publication in the *Proceedings of the American Control Conference*, Baltimore, MD, June 29–July 1, 1994.

[3] J. C. Doyle, K. Glover, P.P. Khargonekar and B.A. Francis, "State-Space Solutions to Standard $\mathcal{H}^2$ and $\mathcal{H}^\infty$ Control Problems," *IEEE Trans AC*, Vol 34, No 8, August 1989.

[4] A.A. Rodriguez and M.A. Dahleh, "$\mathcal{H}^\infty$ Control of Stable Infinite-Dimensional Systems using Finite-Dimensional Techniques," submitted for publication in *IEEE Transactions on Automatic Control*, 1993.

[5] A.A. Rodriguez, "Design of $\mathcal{H}^\infty$ Optimal Finite-Dimensional Controllers for Unstable Infinite-Dimensional Systems," submitted for publication to *AUTOMATICA*, 1993.

[6] A.A. Rodriguez and J.R. Cloutier, "$\mathcal{H}^\infty$ Sensitivity Minimization for Unstable Infinite-Dimensional Plants," *Proceedings of the American Control Conference*, San Francisco, CA, June 2-4, 1993, pp. 2155–2159.

[7] A.A. Rodriguez and M.A. Dahleh, "On the Computation of Induced Norms for Non-Compact Hankel Operators Arising From for Distributed Control Problems," *Systems & Control Letters*, December, 1992.

[8] C.A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, Inc, NY, 1975.

[9] B.A. Francis, *A Course in $H_\infty$ Control Theory*, Springer-Verlag, 1987.

[10] A.A. Rodriguez and J.R. Cloutier, "Control of a Bank-to-Turn-Missile with Saturating Actuators", submitted for publication in the *Proceedings of the 1994 American Control Conference*, Baltimore, MD.

[11] A.A. Rodriguez and J.R. Cloutier, "Control of a Bank-to-Turn-Missile with Multiple Saturating Actuators," in preparation, to be submitted to *AIAA Journal of Guidance, Control, and Dynamics.*

[12] A.A. Rodriguez and S.N. Balakrishnan, "Performance Enhancement for Missile Guidance and Control Systems", proposal submitted for invited session to *1994 American Control Conference*, Baltimore, MD.

[13] S.N. Balakrishnan and A.A. Rodriguez, "Performance Enhancement for Integrated Missile Guidance and Control Systems", Invited Session, *1994 AIAA Guidance and Control Conference*, Phoenix, AZ.

[14] P. Kapasouris, "Design for Performance Enhancement in Feedback Control Systems with Multiple Saturating Nonlinearities," LIDS MIT, PhD Thesis, LIDS-TH-1757, March 1988.

[15] E.G. Gilbert, and K.T. Tan, "Linear Systems with State and Control Constraints: The Theory and Application of Maximal Output Admissible Sets," *IEEE Trans Automatic Control*, Vol AC-36, No. 9, September 1991, pp. 1008–1020.

[16] M. Morari, "Some Control Problems in the Process Industries," Progress in Systems and Control, Essays on Control: Perspectives in the Theory and Its Applications, Birkhauser, Editors: H.L. Trentelman and J.C. Willems, 1993, pp. 55-77.

[17] B.A. Francis and K. Glover, "Bounded Peaking in the Optimal Linear Regulator with Cheap Control," *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 4, August 1978, pp. 608–617.

[18] H.J. Sussman and P.V. Kokotovic, "The Peaking Phenomenon and the Global Stabilization of Nonlinear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-36, No. 4, August 1991, pp. 424–440.

[19] A. Isidori, "Nonlinear Control Systems," 2nd Edition, Springer-Verlag, New York, 1989.

[20] A.A. Rodriguez and Delano Carter, "Hierarchical HAC$_{\mathcal{H}^\infty}$/LAC Vibration Suppression for a Flexible Space Telescope: SPICE," submitted for publication in the *Proceedings of the American Control Conference*, Baltimore, MD, June 29–July 1, 1994.

[21] A.A. Rodriguez and Delano Carter, "$\mathcal{H}^\infty$ Control of SPICE: A Flexible Laser Beam Expander," submitted for publication in the *Journal of Dynamic Systems, Measurements, and Control.*

[22] I.A. Hirsch, M.A. Langehough, J.A. Bossi, et al. , "Advanced Robust Autopilot," Air Force Armament Laboratory, Eglin AFB, Florida, AFATL-TR-89-64, November 1989.

[23] J.H. Blakelock, *Automatic Control of Aircraft and Missiles*, 2nd Edition, John Wiley & Sons, Inc., 1991.

[24] G. Stein and M. Athans, "The LQG/LTR Procedure for Multivariable Feedback Control Design," *IEEE Transactions on Automatic Control*, Vol. AC-32, No. 2, February 1987, pp. 105-114.

[25] T. Kailath, "Linear Systems," Prentice-Hall, 1980.

[26] M. Sonne and A.A. Rodriguez, "PC's in the Design and Evaluation of Guidance and Control Systems for Missiles," to appear in the *Proceedings of the 1994 International Conference on Simulation in Engineering Education*, Tempe, AZ, January 24–26 1994.

[27] M. Sonne and A.A. Rodriguez, "A PC-based Graphics System for the Evaluation of Missile Guidance and Control Laws," submitted for publication in the *Proceedings of the American Control Conference*, Baltimore, MD, June 29–July 1, 1994.

[28] S.C. Warnick and A.A. Rodriguez, "Longitudinal Control of a Platoon of Vehicles with Saturating Nonlinearities," submitted for publication in the *IEEE Transactions on Control Technology.*

## Part I

# Appendix: Proposal for Invited Session to 1994 ACC

A.A. Rodriguez, Chair
Arizona State University
Tempe, AZ 85287-7606

S.N. Balakrishnan, Co-chair
University of Missouri-Rolla
Rolla, MO 65401

## Overview of Proposed Session

Because of the great uncertainty which an evasive target presents, the inherent nonlinear and typically unstable non-minimum phase dynamics associated with missiles, the integrated guidance and control of missile systems still represents one of the richest and most challenging control problems to the systems community. During the past decade many advances have been made in the area of robust and nonlinear control. This includes results which facilitate the design of autopilots, provide insight into gain-scheduling, and allow engineers to better address nonlinear design issues. This session is an effort to bring researchers addressing various aspects of missile guidance and control together to present new results on enhancing the performance of integrated missile guidance and control systems. The session is intended to shed light on some of the issues which researchers in the missile guidance and control community are now addressing.

## Motivation and Summary

The title of the proposed session is

**Performance Enhancement for Missile Guidance and Control Systems.**

To properly address this subject, it is important that all aspects associated with missile guidance and control are addressed. The organizers felt, for example, that it was necessary for the session to contain methods for systematizing, optimizing, and enhancing the performance of today's autopilots and guidance systems. This includes the issue of designing the autopilot to accommodate multiple hard nonlinearities such as saturating actuators, rate limiters, etc. By so doing, autopilot performance would be enhanced and a much more flexible guidance loop could be tolerated. Such a discussion must, however, involve a discussion of various autopilot design methodologies. For this reason, papers will be presented on autopilot design using eigenstructure assignment, $\mathcal{H}^\infty$-based methods, and the latest optimization techniques. The organizers also felt that it was important for the session to contain new methods for improved guidance and target state tracking. This, it was felt could properly be addressed by researchers working on "approximately" analytical guidance laws and nonlinear target state estimation algorithms which exploit the structure of the engagement and the estimator coordinate system.

To accommodate the above requirements, the organizers solicited papers from researchers in academia, industry, and the military. Six (6) papers were selected to address the following topics:

1. **Performance Enhancement for a Missile with Multiple Hard Nonlinearities.**
   A.A. Rodriguez, Arizona State University; Session Organizer.
   James R. Cloutier, Armament Directorate, Eglin Air Force Base.
   title: Control of a Bank–to–Turn (BTT) Missile with Saturating Actuators

2. **Improved Guidance Laws for Missiles: An Analytic Approach.**
   S.N. Balakrishnan, University of Missouri-Rolla; Session Co-organizer.
   title: Improved Guidance Laws for Missiles: An Analytic Approach

3. **Autopilot Design using $\mathcal{H}^\infty$ Design Methods.**
   Kevin A. Wise, McDonnell Douglas Missile Systems Co.
   Eric S. Hamby, McDonnell Douglas Missile Systems Co.
   title: $\mathcal{H}^\infty$ Missile Autopilot Design With and Without Imaginary Axis Zeros

4. **Methods for Improved Target State Estimation.**
   Chris D'Souza, Armament Directorate, Eglin Air Force Base.
   James R. Cloutier, Armament Directorate, Eglin Air Force Base.
   title: Spherical-based Target State Estimation

5. **Autopilot Optimization using Genetic Algorithms.**
   Richard Hull, University of Central Florida.
   Roger W. Johnson, University of Central Florida.
   title: Performance Enhancement of a Missile Autopilot via Genetic Algorithm Optimization Techniques

6. **Missile Eigenstructure Assignment via Dynamic Compensation.**
   Robert Wilson, Armament Directorate, Eglin Air Force Base.
   James R. Cloutier, Armament Directorate, Eglin Air Force Base.
   title: Eigenstructure Assignment via Dynamic Compensation

# ENHANCED LIQUID FUEL ATOMIZATION THROUGH EFFERVESCENT INJECTION

Larry A. Roe
Assistant Professor
Mechanical Engineering Department

Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0238

Final Report for:
Research Initiation Program
Advanced Propulsion Division
Aeropropulsion and Power Directorate
Wright Laboratories

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

December 1993

# ENHANCED LIQUID FUEL ATOMIZATION THROUGH EFFERVESCENT INJECTION

Larry A. Roe
Assistant Professor
Mechanical Engineering Department
Virginia Polytechnic Institute and State University

## Abstract

A single-component, phase-Doppler particle analyzer (PDPA) was assembled, modified for bubble diameter measurements, and applied to several potential configurations of effervescent (2-phase liquid-gas) injection systems. A wide range of instrument hardware and software problems were identified and corrected. The influences of instrument settings and gas loading were evaluated. It was concluded that the PDPA system produces reliable bubble size information for bubble diameters between approximately 2.0 and 800 microns, at gas-to-liquid volume ratios as high as 10 percent (for pressures of approximately 20 psia). These diameters are within the range of interest for effervescent atomization schemes, although the gas loadings are somewhat low. There did not seem to be a strong correlation between bubble size and the size characteristics of the resulting spray. After validation of the measurement concept and preliminary evaluation of experimental hardware designs provided by another contractor, the test system was turned over to USAF and contractor personnel for continued evaluation of injector designs.

# ENHANCED LIQUID FUEL ATOMIZATION THROUGH EFFERVESCENT INJECTION

Larry A. Roe

## INTRODUCTION

Effervescent injection offers the potential for significant performance improvements in all liquid-fueled propulsion systems, with advantages particularly suited to ramjet engines. This fuel injection scheme would typically introduce small bubbles of air or another gas into the liquid fuel stream prior to injection into the combustion chamber. The bursting of these bubbles leads to rapid breakup of the liquid fuel and dramatically improved atomization at low fuel-supply pressures. A major drawback to the successful implementation of this technique in actual combustion systems is the total lack of numerical data relating bubble size in the two-phase fuel stream to the diameter of the resultant fuel spray droplets. A primary reason for this shortcoming has been the unavailability of appropriate instrumentation capable of measuring the bubble size distributions.

One difficulty typically encountered when evaluating the performance of liquid fuel injectors is associated with the acquisition of reliable droplet and bubble statistics. Parameters crucial to the characterization of the injection scheme include average droplet diameter, droplet size distribution, droplet velocity, and diameter-velocity correlations. In addition, any attempt to correlate droplet statistics with bubble statistics in 2-phase injectors requires a reliable technique for bubble sizing. The focus of this program was to provide the instrumentation and analysis capability required for such evaluations. The instrumentation system, a phase-Doppler particle analyzer (PDPA), provides an analysis capability for effervescent injection studies which has not been previously utilized by researchers in this field.

## REVIEW: TWO-PHASE INJECTION

One of the earliest investigations of the atomization of a gas-liquid mixture was reported by Chawla (1985). It was determined that small droplets were produced, largely independent of the size of the fuel delivery orifice. In addition, small droplets were produced at relatively low fuel velocities when compared to pressure atomizers.

The concepts associated with effervescent injection have been developed in papers by Wang et al. (1987), Roesler and Lefebvre (1987), Lefebvre et al. (1988), Avrashkov et al. (1990), and Arai and Schetz (1992).

Wang et al. (1987) studied effervescent injection of water/nitrogen mixtures into quiescent air at normal atmospheric pressure and temperature. The nitrogen was bubbled into a large chamber immediately upstream of the injection orifice. Two gas injection designs (differing primarily in the diameter of the holes through which the gas was bubbled into the liquid) and three orifice diameters were evaluated. Gas pressure and gas/liquid mass fraction were each varied over about an order of magnitude. Droplet size was determined with a Malvern analyzer, which provides a spatially averaged mean droplet size. Bubble sizing was not attempted. It was concluded that the atomization varied primarily with injection pressure and mass ratio, with less sensitivity to orifice diameter and injection geometry.

Roesler and Lefebvre (1987) extended the previous study to a wider range of gas/liquid mass ratios. Air was introduced through a porous cylinder into the water stream, and the influence of air supply pressure, aerator tube porosity, orifice diameter, and gas/liquid mass ratio were evaluated. Again, orifice size and aerator porosity (which was assumed to control bubble size) were found to have little influence on the mean droplet size. Bubble sizes were not measured. Good atomization occurred for gas/liquid mass ratios as low as 0.02. The air pressures required were only those sufficient to cause flow through the porous cylinder at the mass rates desired.

A continuation study by Lefebvre et al. (1988) confirmed these conclusions over a different range of conditions. Good atomization was again obtained using small amounts of injected air, at injection pressures as low as 5 psi. Avrashkov et al. (1990) used both hydrogen and helium bubbles in kerosene, with gas/liquid mass ratios from 0 to 10 percent. The mixture was injected at high pressure (~ 20 atm) into a supersonic combustion chamber at 1 atmosphere and the self-ignition and stability characteristics evaluated. The gas addition was found to increase the dispersion of the spray cone, provide better liquid penetration into the free stream and improve mixing, but did not significantly affect mean drop size at the high injection pressures utilized.

Arai and Schetz (1992) injected helium/water flows at high pressure (~20 atm) into a supersonic tunnel through an array of 0.8 mm diameter orifices. The production of the small bubbles required to maintain bubbly flow through such small passages required the addition of a surfactant to the liquid prior to gas injection. Photographic analysis of the resulting sprays showed that the gas injection increased the plume dispersion angle and increased penetration for single-orifice injection. Changes in surfactant concentration were found to affect the spray characteristics, apparently due to changes in bubble size, although this was not a measured parameter.

In addition to the published information summarized above, more recent, but as yet unpublished, research efforts are in process to evaluate the application of effervescent injection to scramjet engines (Northam 1992). As with the prior studies, the porosity of the air injection cylinder was found to have little effect on the resulting spray characteristics.

## DESCRIPTION OF PHASE-DOPPLER PARTICLE ANALYZER

The PDPA system as manufactured by Aerometrics is based on a development by Bachalo and Houser (1984). A more thorough description of the application of this technique is provided by Bachalo et al. (1991). A complete description of the operational principles is well beyond the scope of this report, but a brief summary is appropriate.

A PDPA is essentially a single-component laser Doppler anemometer (LDA) with multiple photodetectors and additional signal processing capability. As an LDA system, it is of the standard dual-beam type. Two laser beams intersect at a small angle in the region where measurements are to be obtained. The crossing of these two beams defines a probe volume; droplets or bubbles passing through this zone scatter light simultaneously from the two beams. When this scattered light reaches the receiving optics (lens or photodetector), it forms an interference pattern, which moves in space due to the droplet motion. The temporal frequency at which the interference fringes sweep across the surface of the detector is related to the transmitting optics (laser wavelength and beam intersection angle) and the particle velocity component in the plane of the intersecting beams. Since the optical configuration is known, the velocity component can be determined.

Additional information is required for sizing. This is provided by additional photodetectors. Essentially, these photodetectors image different regions of the interference pattern as it sweeps across the receiving optics. The photodetectors all observe the same temporal frequency (the Doppler frequency related to velocity) but observe different spatial frequencies since the interference pattern fringes are not parallel. This translates to a phase difference between detectors. This phase difference is related to the curvature of the fringe pattern, which is related to droplet or bubble diameter (and a long list of other parameters, which are generally known). With sophisticated signal processing and data analysis, the diameter (assumed spherical) can be determined.

An earlier version of the PDPA system had been fully characterized by the Principal Investigator during the 1992 Summer Faculty Research Program and utilized for droplet measurements (Roe 1992). The primary instrument operating parameters which were found to influence the results were; the incident beam intersection angle, frequency shifting of the incident beams by a rotating diffraction grating, receiving optics alignment, photomultiplier (PMT) voltage, and filtering of the output signal. Several modifications to the system hardware and software occurred in the time period between the Summer 1992 effort and the beginning of this research program. In addition, further modifications to the operation of the system were required for application to bubble sizing.

## EXPERIMENTAL PROGRAM AND RESULTS

The experimental program was conducted on-site at the laboratories of the Advanced Propulsion Division, Aeropropulsion and Power Directorate, Wright Laboratories, Wright-Patterson AFB, Ohio. The program had two aspects. Primarily, the PDPA system was modified to provide reliable operation for bubble sizing measurements, debugged, characterized, and made operational. Secondarily, experimental apparatus designed and constructed in conjunction with another on-site contractor was evaluated, using the PDPA as the primary analysis tool.

Bubble Generators

Personnel from CFD Research provided support for the design and construction of several bubble generating systems which produced two-phase flows suitable for effervescent atomization. Main operational criteria included the ability to reliably produce bubbles with specific size distribution characteristics, optical access for the PDPA system, controllability, and repeatability. This task proved more difficult than originally anticipated.

The bubble generation was accomplished by bubbling air into a flowing stream of water. Three different configurations were tested. The first utilized an existing 1-inch square test section, with a two-dimensional, variable area, converging-diverging section to control bubble generation. Maximum use of pre-existing hardware was made to minimize long lead times in the laboratory machine shops. Water flowed through the duct, and a tube with small holes was inserted such that the air bubbled through the holes into the water. The position of the tube could be varied throughout the variable area section of the assembly, so that the local water velocity at the injection point would vary. It was hoped that this would alter the shearing action of the water on the bubble ports and change bubble size controllably. Bubble size could, in fact, be varied, but repeatability was not acceptable. Additionally, optical access required for good PDPA measurements was not sufficient.

The second bubble generator utilized a 1-inch diameter glass tube for maximum optical access, and two types of air injection schemes. The water flowed through an axisymmetric converging-diverging nozzle, with air injection either through a centerline tube, as with the first configuration, or directly through the wall at the nozzle throat. The majority of the PDPA evaluation was conducted with this generator. Controllability of the bubble generation process was still deficient, but, as the primary goal of the program was the establishment of the PDPA as a measurement technique, this generator proved sufficient to achieve that end.

A third configuration was assembled by CFD personnel, and some preliminary testing was done. This design featured a porous plate bubble generator, multiple water inlets, swirling flow, accumulator tanks, and surge eliminators. Continued modification and testing of this device is still in progress.

PDPA Evaluation

The primary goal of the program was to establish the capability of utilizing phase-Doppler particle analysis as the primary evaluation tool for 2-phase, liquid-gas injection schemes. Although the instrumentation had been previously utilized for droplet measurements (Roe 1992) several modifications to instrument setup and operation were required for bubble sizing.

Initial testing of the instrument was conducted with a conventional optical alignment, 30-degree forward scatter light collection. Major difficulties were encountered, including poor signal-to-noise ratio, low data rates, obviously incorrect sizing data, and non-repeatable results. The instrument configuration was modified to collect scattered light in the 63-degree off-axis back scatter direction, but difficulties continued. An extended assessment followed, starting with the system optics, electronics, and processors. It was eventually discovered that a factory-supplied system software upgrade loaded prior to the start of this research program was faulty. In this system, the software controls the entire data acquisition and reduction, including varying the spacing between transmitted beams, controlling frequency shift settings, setting filter cutoffs in the electronics, and calculating diameters and velocities from the measured frequency data. The parameters loaded with the software upgrade were, in fact, not completely consistent with the hardware. This lead to a wide range of perplexing difficulties with the system. Eventually, portions of the previous version software were located and loaded, leading to acceptable system operation.

The system was validated and exercised on the second-configuration CFD bubble generator. The system was concluded to provide acceptable operation for bubble diameters between approximately 2 to 800 microns, for gas-to-liquid volume ratios as high as 10 percent. This corresponds to a mass ratio of approximately 0.5 percent at a pressure of 20 psia. Effervescent injection was demonstrated at this gas loading, but typical applications are anticipated to have higher gas-to-liquid ratios. The limitation on gas loading is primarily the obscuration of the measurement region by bubbles between the measurement region and the optics. It may prove feasible to obtain good PDPA data at higher loadings by passing the flow, or some portion of it, through an optically thin section between two glass plates, so that an essentially 2-dimensional slice of the flow can be examined.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to all Advanced Propulsion Division personnel who contributed to this research effort. Special appreciation goes to Kevin Kirkendall of CFD Research for continued and valiant engineering support in the design, fabrication, assembly, and testing of the bubble generation systems, and to Alan Spring of CFD for guidance in the conceptual design and application of effervescent injection systems. The technical guidance and support of Abdi Nejad are especially appreciated.

## REFERENCES

Arai, T., and J. A. Schetz, "Penetration and Mixing of Bubbling Liquid Jets From Multiple Injectors Normal to a Supersonic Air Stream," submitted for publication to AIAA, Oct. 1992.

Avrashkov, V., S. Baranovsky, and V. Levin, "Gasdynamic Features of Supersonic Kerosene Combustion in a Model Combustion Chamber," AIAA paper 90-5268, AIAA Second International Aerospace Planes Conference, Orlando, Oct. 29-31, 1990.

Bachalo, W. D., and M. J. Houser, "Phase/Doppler Spray Analyzer for Simultaneous Measurements of Drop Size and Velocity Distributions," Optical Engineering, vol 23, no 6, pp. 583-590, Sept/Oct 1984.

Bachalo, W. D., A. Brena de la Rosa, and R. V. Sankar, "Diagnostics for Fuel Spray Characterization," Combustion Measurements, N. Chigier, ed., pp. 229-278, Hemisphere, 1991.

Chawla, J. B., "Atomization of Liquids Employing the Low Sonic Velocity of Liquid/Gas Mixtures," Proceedings of the 3rd International Conference on Liquid Atomisation and Spray Systems, London, 1985.

Lefebvre, A. H., X. F. Wang, and C. A. Martin, "Spray Characteristics of Aerated-Liquid Pressure Atomizers," Journal of Propulsion, vol. 4, no. 4, pp. 293-298, July-Aug. 1988.

Northam, G. B., personal communication, Oct. 1992.

Roe, L. A., "Determination of the Operational Characteristics of a Phase-Doppler Droplet Analyzer and Application to a Ramjet Fuel-Injection Research Tunnel," final report for AFOSR Summer Research Program, Sept. 1992.

Roesler, T. C., and A. H. Lefebvre, "Studies on Aerated-Liquid Atomization," ASME paper 87-WA/HT-17, Winter Annual Meeting, 1987.

Wang, X. F., J. S. Chin, and A. H. Lefebvre, "Influence of Gas-Injector Geometry on Atomization Performance of Aerated-Liquid Nozzles," 24th National Heat Transfer Conference, ASME HTD vol. 74, pp. 11-18, 1987.

# Sensor Data Clustering and Fusion for IR/MMW Dual-Mode Sensors Using Artificial Neural Networks

Thaddeus A. Roppel
Associate Professor
Department of Electrical Engineering

Auburn University
200 Broun Hall
Auburn, AL 36849-5201

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.
and
Auburn University

December 1993

# Sensor Data Clustering and Fusion for IR/MMW Dual-Mode Sensors Using Artificial Neural Networks

Thaddeus A. Roppel
Associate Professor
Department of Electrical Engineering
Auburn University

## Abstract

Experiments with artificial neural network processing of infrared imagery have shown that considerable performance improvement can be obtained if the available data are clustered before training. The data are clustered into two categories: normal and outlier. These categories are differentiated by the average and standard deviation of the RMS error that results when a neural network is trained using the leave-one-out method applied to a specific image. Applying the clustering algorithm reported here, we find that performance is improved from 30% to 80% correct classification for 10 x 10 pixel images, while the incorrect classification rate simultaneously drops from 30% to 0%. For 20 x 20 pixel images, correct classification improves from 29% to 62%, while incorrect classification drops from 54% to 12%. The results for fusion of 10 x 10 images with 20 x 20 images are similar to those for 20 x 20 images alone.

# Sensor Data Clustering and Fusion for IR/MMW Dual-Mode Sensors Using Artificial Neural Networks

Thaddeus A. Roppel

## INTRODUCTION

This report includes work done from 1 January to 31 December 1993 under the Air Force Summer Research Extension Program. The laboratory focal point for this work was Mr. Ellis Boudreaux, WL/MNG-X. This work would not have been accomplished without the key contributions made by Mrs. Mary Lou Padgett, Auburn University Research Associate, by Mr. Mark Townsley, and by Graduate Researchers Mr. Camille Raad and Mr. Tobias Graf von Haslingen.

Sensor fusion has the potential for improving detection and identification of targets. According to recent studies, both theoretical and experimental, the amount of improvement can range from almost zero dB to well over 3 dB, depending on the signal-to-noise ratio (SNR) of each sensor, and the sensor noise correlation. The improvement can be effectively infinite in the case where one sensor fails completely. Neural networks have been under consideration for raw data fusion due to the high speed of response needed and the complexity of the problem. This study addressed certain specific questions regarding the training of a neural network to accomplish raw data fusion under conditions where sparse data are available. The objective is to obtain generalizable results from a small data set with unequally sampled categories. An algorithm is suggested by which a neural network can be trained to take maximum advantage of existing data, as well as incorporating new data to maximum advantage.

## METHODOLOGY

In this work, artificial neural networks (ANN's) are trained on IR images of three types of ground targets. T-62 tanks, M-113 APC's, and Lance missile launchers. The available data set contains 10 tank images, 8 APC's, and 4 launchers. Each image is a 40 x 40 pixel image with 8-bit gray-scale pixels and is averaged down to 10 x 10 pixels using a neighborhood averaging technique. The neural networks were simulated using the Aspirin/MIGRAINES (A/M) package developed by Mitre [Leighton, 93]. The architecture / training is feedforward with classical backpropagation of errors with the number of input nodes equal to the number of pixels, 12 hidden layer nodes, and two or three output nodes (one per target category).

A set of analysis tools developed by Mr. Mark Townsley and enhanced by Mr. Camille Raad, both Auburn University Research Assistants, forms the basis for the results generated. These tools supplement the neural network to assist in determining clusters within target type which have similar average RMS error per image. The interactions of the images of different categories are observed from graphs of successful identification rate *versus* the decision threshold [Masters, 93; Padgett et al. 93]. Much of the work illustrated is an extension of the techniques described in [Webster et al.84; Padgett et al.85; Padgett, Roppel 92 ; and Padgett, Karplus 84] and using the NASA Nets multilayer perceptron simulator described in [Savely et al. 90].

The motivation for the current approach came from several sources. Exploration of the A/M simulator capabilities led to investigation of its "rocks and mines" example, which is similar in principle to our tank/launcher/APC problem, and to conversations with T. Sejnowski [Gorman and Sejnowski 88]. Alexis Weiland, of UCLA, also discussed the use of the Principal Components Analysis (PCA) and Canonical Discriminant Analysis (CDA) tools in MIGRAINES. Further insight was provided by [Dai 92 and Gluck et al. 92].

In the example using sonar data to classify objects as rocks or mines, a large training set was available, and many angles of incidence were included in this set. Being able to control the sample data, the environment and replications allowed the researchers to control the variance and sample size. However, as in many real-world situations, the project reported here is expected to use hard-to-obtain, expensive images with unequal sample sizes, faulty sensors, and few target rotations. The results of individual flight tests are unequally represented in the training images, and the range of flight conditions is limited compared to those expected in actual combat. It is also a condition of this study that the image data be presented to the neural network with as little pre-processing as possible. Any information generated about the images is considered to be valuable for the design of future data collection experiments and for comparing data fusion techniques.

Future enhancements of the neural network strategy are targeted toward principles discussed by Sejnowski as extensions of his earlier work. Colleagues of M. Arbib are also investigating the fusion of multisensory input using feedforward neural networks with backpropagation training and limited connectivity [Fagg and Arbib 92]. Essentially, if two banks of sensors are providing input, the hidden layer nodes may be divided into regions. There may be regions accepting input from only one sensor bank type and a region accepting input from both types to illustrate their interaction. The degree of overlap is controllable and an excellent experimental tool. The separation needed for identification of different features found in the input sets might require the addition of another hidden layer. Such architectural modifications to the current single hidden layer, fully connected neural network are potentially useful, but will not be employed unless deemed necessary with the addition of more image types and sensor types in the future.

For the current set of IR images, experts cannot routinely visually detect the target category (tank, APC or launcher). Some clustering of images can be visually detected, but before analysis it is difficult to decide the significance of this clustering. The methodology presented below extends that described in [Padgett et al 93].

First, the average RMS error per image is computed using the entire set of images and jackknifing using the leave-one-out method [Masters, 93 p. 12]. All but one example from each pattern category (categories in this case include tanks, APC's, and missile launchers) is included in the training set, and the left-out examples from each category are used for testing. This procedure is repeated until each image has been used as a test image for an artificial neural network. Ideally, if the training sets for each category are homogeneous in some sense detected by the neural network, the RMS error of each image in a category will be close to the average RMS error for that category [Savely, et al. 1990]. In a well-trained neural network, the average and maximum RMS errors should be low for the images in the training set and for the image(s) in the testing set.

For a second determining factor, including the interactions among images, the confusion matrix results versus threshold are diagrammed. The RMS error generated by each test set of images (one from each pattern category) is graphically illustrated by grouping the results into correct, incorrect and ambiguous responses based on a threshold. This threshold is intended to be varied by the user according to the penalties associated with incorrect results (e.g. hitting a manned vehicle), versus setting a flag indicating an "uncertain" condition to be dealt with in an increasingly vigilant manner in the next set of images in the series.

The leave-one-out results and the confusion matrix results are used to cluster the available data into "normal data" and "outlier data." The normal data are the images which are self-consistent and mutually reinforcing. They will lead to good neural network performance in the sense that the network will correctly classify them with a high probability regardless of which images are used for training or testing. The outlier data are those images which tend not to be recognized by the network trained on the normal images. In other words, the network has trouble generalizing from the normal images to the outlier images. This can occur due to a number of conditions. For example, an outlier image may have been collected under extreme illumination conditions, or perhaps it contains excessive noise or obstruction.

We are presently in the process of generating an objective, computer-based fuzzy logic algorithm to perform the clustering discussed above, but presently it is done manually. For this analysis, the magnitude of the standard deviation is considered along with the average RMS error. These results are used to separate the images of each target type into two clusters, "normal" and "outlier."

## RESULTS

The results are presented graphically, accompanied by discussion. There are two types of graphs, each of which requires some explanation before it can be usefully read.

### Explanation of Graph Formats

Graph Type 1: RMS Error Results versus Images (e.g., Figure 1, top). On this graph type, each data point on the horizontal axis represents an individual image from the available data set. The images are grouped by type (tank, APC, or launcher), and within each type the images are numbered starting from zero. The numbering is consistent from graph to graph, so that "Launcher 0" always refers to the same image, etc. The vertical axis scale is the percent RMS error. This is the neural network error, as measured by the RMS deviation of the output from the ideal values of 1.0 or 0.0. Associated with each image are three data point symbols; two open circles and an asterisk (*). The asterisk marks the value of the average RMS error, as accumulated over all the leave-one-out runs where this image was left out and then trained on. The open circles mark the average plus and minus twice the standard deviation of the error. Inclusion of both open circles is intentionally redundant, but leads to the disconcerting appearance of a negative RMS error, which is an artifact. On this type of graph, outlier images tend to appear as having high average RMS error and/or large standard deviation of error.

Graph Type 2: Confusion Matrix Results versus Threshold (e.g., Figure 1, bottom). On this graph type, the images are not individually represented. The vertical axis measures the percent of the total number of leave-one-out trials. The total number of trials for a particular graph depends on which images are included in the experiments. For example, if launchers and tanks are included, then the total number of trials will be (10 tanks) x (4 launchers) = 40 trials. The horizontal axis, labeled "Threshold," is the decision threshold for the confusion matrix. Each time an image is tested, the neural network outputs take on values from zero to one (actually, 0.05 to 0.95 due to simulator convergence requirements). If a particular threshold value is chosen, then the outputs can interpreted thus: target correctly identified, target incorrectly identified, or indeterminate / unclassified. The latter case occurs when the outputs are between 0.05 + threshold and 0.95 - threshold. The higher the threshold value, the less likely an indeterminate value will be obtained, since more results are forced to be considered either correct or incorrect. A perfectly performing neural network would show 100 percent correct classification at the lowest value of threshold. Each data point (value of threshold) has three data markers associated with it. An asterisk is used to indicate the percentage of correct classifications, an open circle indicates the percentage of indeterminate classifications, and an "X" is used to indicate the percentage of incorrect classifications.

## Graphical Results and Discussion

Figure 1 shows the results of using all the available data to train the neural network, without trying to identify outliers. Only about 30% of the targets are correctly identified at the maximum threshold level, while 30% are incorrectly identified and 40% are indeterminate. Figure 2 shows the substantial improvement that results from identifying the outliers and forcing them into the training set at all times. In this case the correct identification is made about 80% of the time, and no incorrect identifications are made. The improved performance is also evident from the graph of RMS error results, which shows reduced average RMS error and standard deviations. The numerical results from these two figures are summarized in Table 1. Also shown in Table 1 are the confusion matrix results for the two-target experiments. Figure 3 shows the graphical results for tanks and APC's, Figure 4 shows APC's and launchers, and Figure 5 shows tanks and launchers. In each figure, the outliers are identified from the RMS error graphs as those images having the largest average RMS error and/or the largest standard deviations. After identifying all of the outliers, the outliers are forced into the training set, with the results as described above, in Figure 2, and in the last row of Table 1.

Table 1. Confusion matrix results for various training conditions. All data are read from the indicated figures at maximum threshold.

| Targets | Correct % | Incorrect % | Unclass. % | Outliers Detected | Comments |
|---|---|---|---|---|---|
| Tanks<br>APC's<br>Launchers | 30 | 30 | 40 | Launcher 0<br>APC 1, 5, 6<br>Tank 7 | All data treated equally. Figure 1. |
| Tanks<br>APC's | 82 | 14 | 4 | APC 1<br>Tank 7 | All data treated equally. Figure 3. |
| APC's<br>Launchers | 53 | 28 | 19 | Launcher 0<br>APC 0, 1 | All data treated equally. Figure 4. |
| Tanks<br>Launchers | 48 | 30 | 22 | Launchers 0,1<br>Tank 3 | All data treated equally. Figure 5. |
| Tanks<br>APC's<br>Launchers | 80 | 0 | 20 |  | Outliers forced into training set. Figure 2. |

Visual inspection of the final clustering of the images revealed a solid technical basis for the need to include these images in the training set. Only with the high-RMS error images included in the training set were the complete set of available flights sampled. Due to cloud conditions and time of day, image intensity variation between flights was significant. These results suggest that some effort needs to be directed at reducing the intensity sensitivity of the neural network.

Image Resolution Experiments

In order to further investigate the IR imagery, we conducted a series of experiments using higher resolution images. The experiments described previously used 40 x 40 pixel images reduced to 10 x 10 pixels by neighborhood averaging. We generated a set of IR images starting from the same 40 x 40 arrays, but averaging down only by a factor two in each dimension to 20 x 20 pixels. These higher resolution images were used in the same type of experiments as described previously. The neural network architecture was slightly modified to accommodate the increased input array. In order to test the effect of resolution on identification, a series of experiments was conducted with the 20 x 20 images. We hypothesized that increased resolution would provide better target discrimination since more information was available to the neural network. We were somewhat surprised, therefore, to find that the performance was considerably worsened. The results are shown in Figure 6, and may be compared with those shown in Figure 1. Notable improvement was achieved, as in the 10 x 10 case, by forcing the outliers into the training set. These results are shown in Figure 7, which may be compared with Figure 2. These comparisons are summarized in Table 2. These results suggest that further work is required to optimize the neural network architecture in order to accommodate the increased resolution images.

Table 2. Comparison of 10 x 10 performance with 20 x 20 performance at maximum threshold.
LEGEND: "*" = percentage correct, "X" = percentage incorrect, "O" = percentage unclassified.

| Image Resolution | All images tested | Outliers forced into training set |
|---|---|---|
| Tanks, APC's, Launchers<br>10 x 10 pixels | * = 30%<br>X = 30%<br>O = 40% | * = 80%<br>X =  0%<br>O = 20% |
| Tanks, APC's, Launchers<br>20 x 20 pixels | * = 29%<br>X = 54%<br>O = 17% | * = 62%<br>X = 12%<br>O = 26% |

Sensor Fusion Experiments

Our original objective in this project was to investigate IR / MMW radar sensor fusion. However, there has been an extensive delay in obtaining valid radar data. As of this writing, we have just been able to extract radar data from a data set that should prove useful in future experiments. Nevertheless, we put considerable effort into designing the fusion experiments with the goal of being ready to conduct the experiments as quickly and efficiently as possible when the data are ready. In order to test our algorithms, and to further investigate the high resolution results reported above, we conducted a series of experiments on fusion of the low resolution images (10 x 10) with the higher resolution images (20 x 20). Our hypothesis was that some improvement should result, since the low resolution data would provide coarse feature information, while the higher resolution would provide differentiating details. The results are shown graphically in Figure 8 (all images treated equally) and in Figure 9 (outliers forced into training set). The results at maximum threshold are summarized in Table 3. From this data, it is evident that fusion enhancement is not significant for the experimental conditions shown here. The fusion results are almost identical to the results for 20 x 20 alone, which seems to support the idea that the neural network is not optimized for the higher resolution images.

Table 3. Low resolution, higher resolution, and fusion performance at maximum threshold.

| | All images tested | | | Outliers forced into training set | | |
|---|---|---|---|---|---|---|
| Image Resolution | Correct % | Incorrect % | Unclass. % | Correct % | Incorrect % | Unclass. % |
| 10 x 10 | 30 | 30 | 40 | 80 | 0 | 20 |
| 20 x 20 | 29 | 54 | 17 | 62 | 12 | 26 |
| FUSION | 24 | 53 | 23 | 61 | 11 | 28 |

Dual Mode Data Extraction from LOSA

The goal of this subtask is to use existing software and data to establish a dual mode image set for sensor fusion experiments at Auburn. This subtask was begun during the author's 1993 summer research program at Wright Laboratory, and continued upon returning to Auburn. The data set consists of co-boresighted MMW / IR data taken on a low-observable subsonic aircraft (LOSA). The packed raw data together with a large ground-processing software (GPS) package was made available to us. The original GPS was written for a specific VMS-based computer system A considerable amount of time has been spent modifying the GPS code to run under UNIX in a more flexible and portable environment, since Auburn no longer has a supported VMS system. As of this writing, we have been able to view IR still frames from the packed data, and also we have been able to view coarse-range gate MTI maps that are time-tagged to the IR images. We will begin fusion experiments immediately with this data, however, time will not permit the results to be included in this report. We intend to report the results in a briefing to WL/MN during January 1994.

## CONCLUSIONS

The most significant result of the experiments conducted under this contract is the development of an algorithmic method to separate the available data into distinct categories for training purposes. For best performance, it is essential to know which data will enhance performance and which will be confusing to a particular neural network architecture. Principal components analysis (PCA) and canonical discriminant analysis (CDA) are conventional techniques which have the same objective. It is worthwhile to investigate the combined use of these tools with the algorithm we have reported here. We have been successful in installing the LOSA data and ground processing software, and are about to begin IR / MMW sensor fusion experiments.

## REFERENCES

Dai, Han-Sen. 1992. System Identification Using Neural Networks. Ph. D. Dissertation. UCLA. Los Angeles., CA.

Fagg, A. H. and M. A. Arbib. 1992. "A Model of Primate Visual-Motor Conditional Learning." Journal of Adaptive Behavior, Summer, 1992.

Gluck, R. H.-S. Dai and W. J. Karplus. 1992. "On-Orbit Nonlinear Structural Parameters Realization via Artificial Neural Networks." AIAA/ ASME / ASCE / AHS /ASC 33rd Structures, Structural Dynamics and Materials Conference. April, 1992. Dallas, TX.

Gorman, R. P. and T. J. Sejnowski. 1988. "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets." Neural Networks. Vol. 1, 75-89.

Leighton, Russell, Aspirin/MIGRAINES, Version 6, 1993, Mitre Signal Processing Center, McLean, VA.

Masters, Timothy. 1993. Practical Neural Network Recipes in C++. Academic Press, Inc.

Padgett, Mary Lou and Walter J. Karplus. 1984. "Computational Intelligence Standards: Motivation, Current Activities and Progress." Computer Standards and Interfaces, July-August, 1994 (invited, in press).

Padgett, Mary Lou, S. A. Rajala, W. E. Snyder, and W. Howard Ruedger. 1985. "Detection of Maneuvering Target Tracks." Proceedings SPIE 29th Annual In. Tech. Sym. on Optical and Electro-optical Engineering. San Diego, August 18-23, 1985.

Padgett, Mary Lou and T. A. Roppel. 1992. "Neural Networks and Simulation: Modeling for Applications." Simulation. Vol. 58: No. 5, May, 1992. pp. 295-305.

Padgett, Mary Lou , T. A. Roppel, C. C. Raad, M. Townsley and T. Graf von Haslingen. 1993. "Neural Networks for Signal Processing and Analysis: A Clustering Approach." Proceedings of the Fifth Workshop on Neural Networks: AIND. (San Francisco, CA), Nov. 1993. pp. 384-89.

Savely, R., R. Lea, R. Shelton, J. Villareal, and L. Wang. 1990. "Overview of Advanced Architectures Research and Development Activities in the Software Technology Branch." Proc. First Workshop on Neural Networks: AIND (Auburn, AL Feb. 1990). pp. 3-14.

Webster, D. B., M. L. Padgett, G. S. Hines , D.L. Sirois. 1984. "Determining the Level of Detail in a Simulation Model -- A Case Study." Computers in Industrial Engineering. Vol 8. No 3/4. Dec. 1984. pp. 215-255.

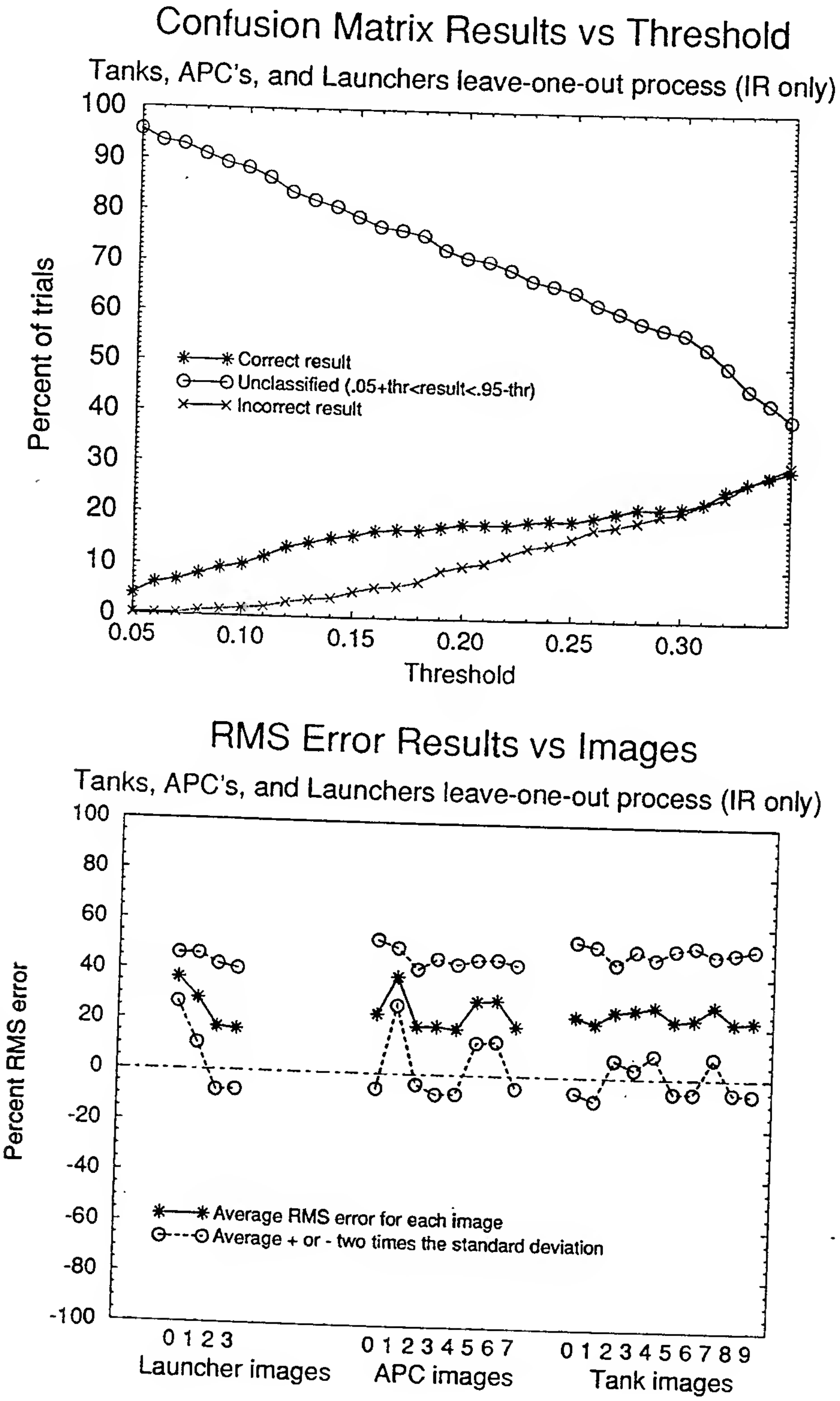


Figure 1. Neural network performance on all three target types with no clustering of images.

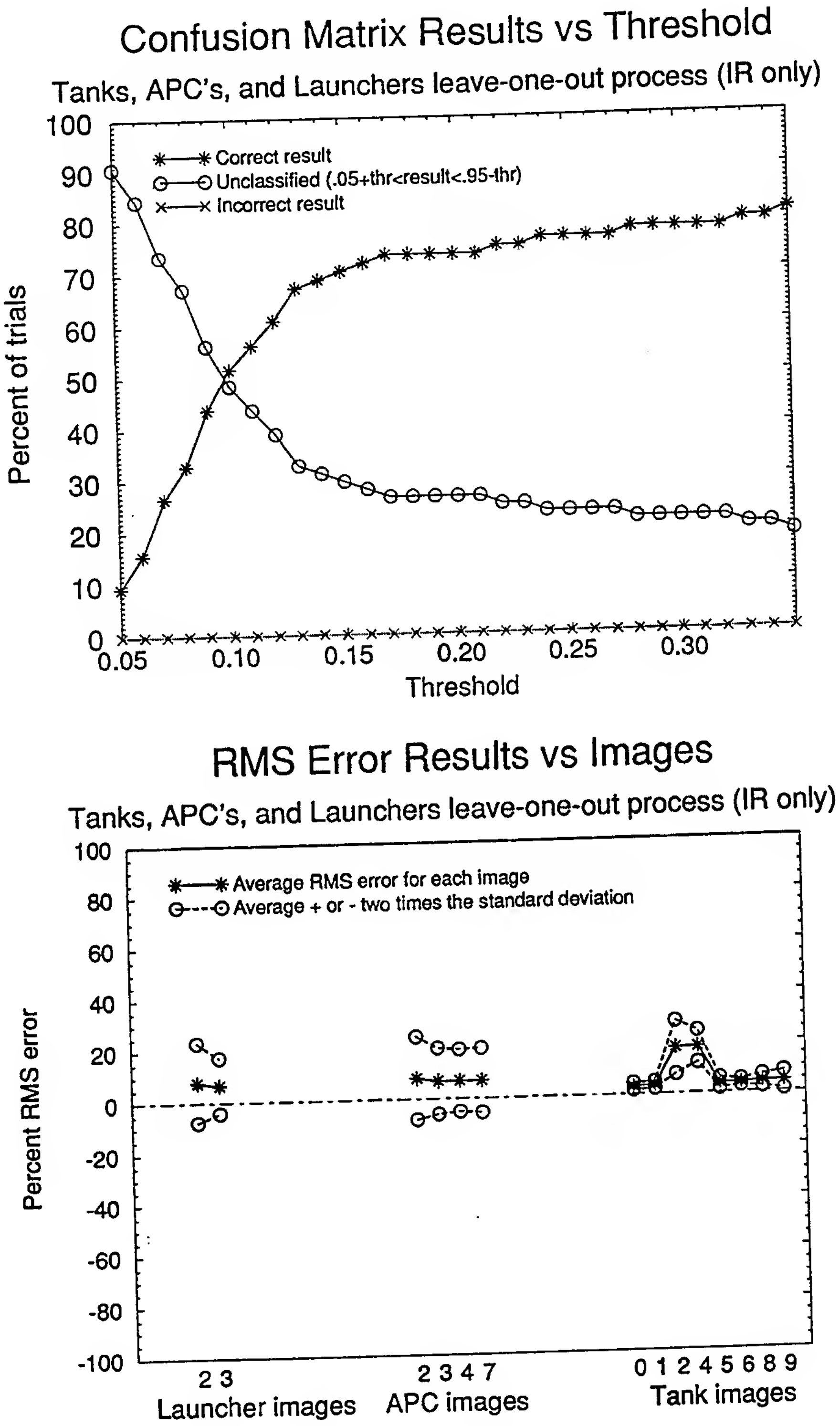


Figure 2. Neural network performance on all three target types after applying the image clustering algorithm.

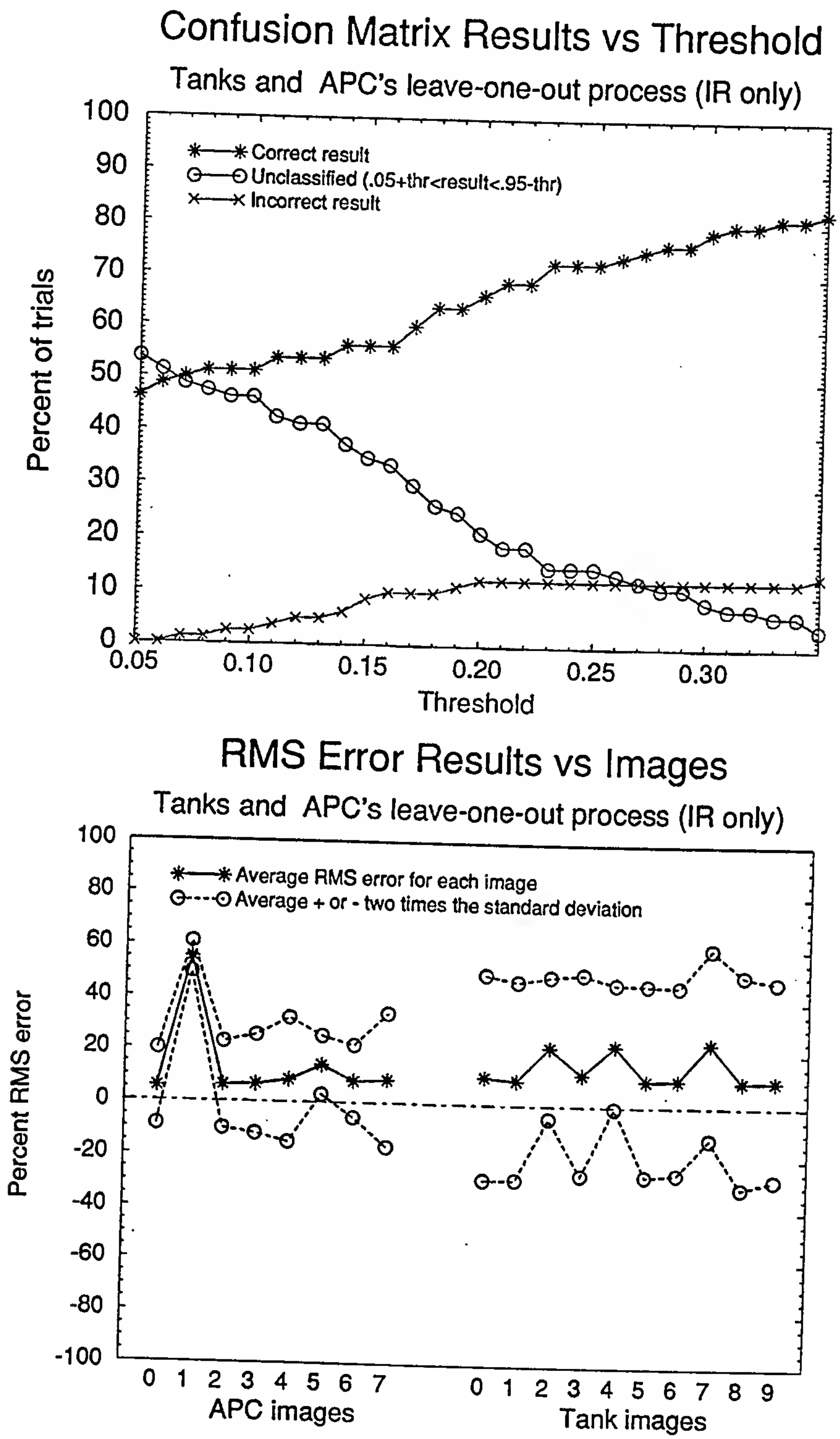
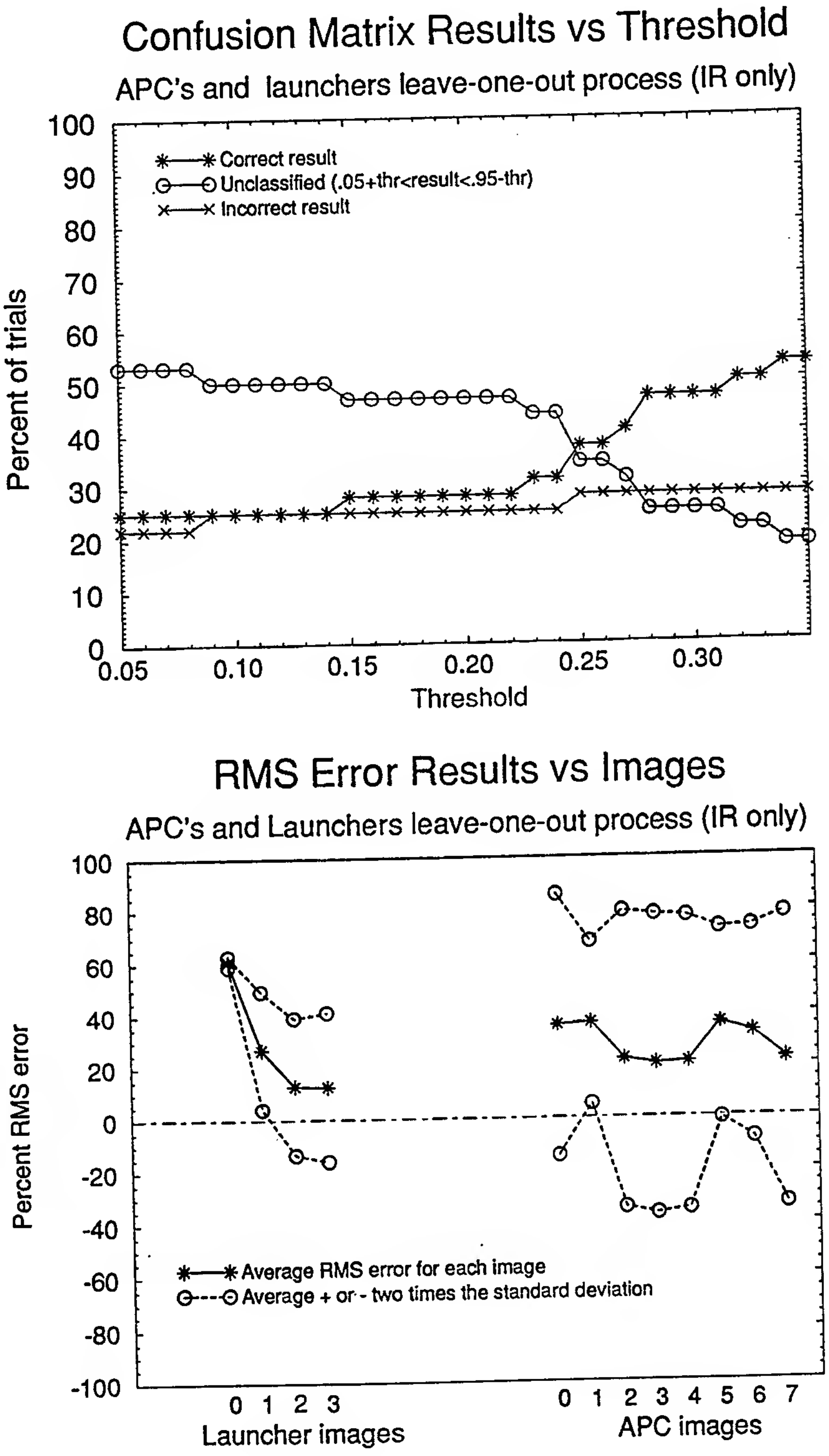


Figure 3. Neural network performance on two target types with no image clustering.



Confusion Matrix Results vs Threshold

APC's and launchers leave-one-out process (IR only)

RMS Error Results vs Images

APC's and Launchers leave-one-out process (IR only)

Figure 4. Neural network performance on two target types with no image clustering.

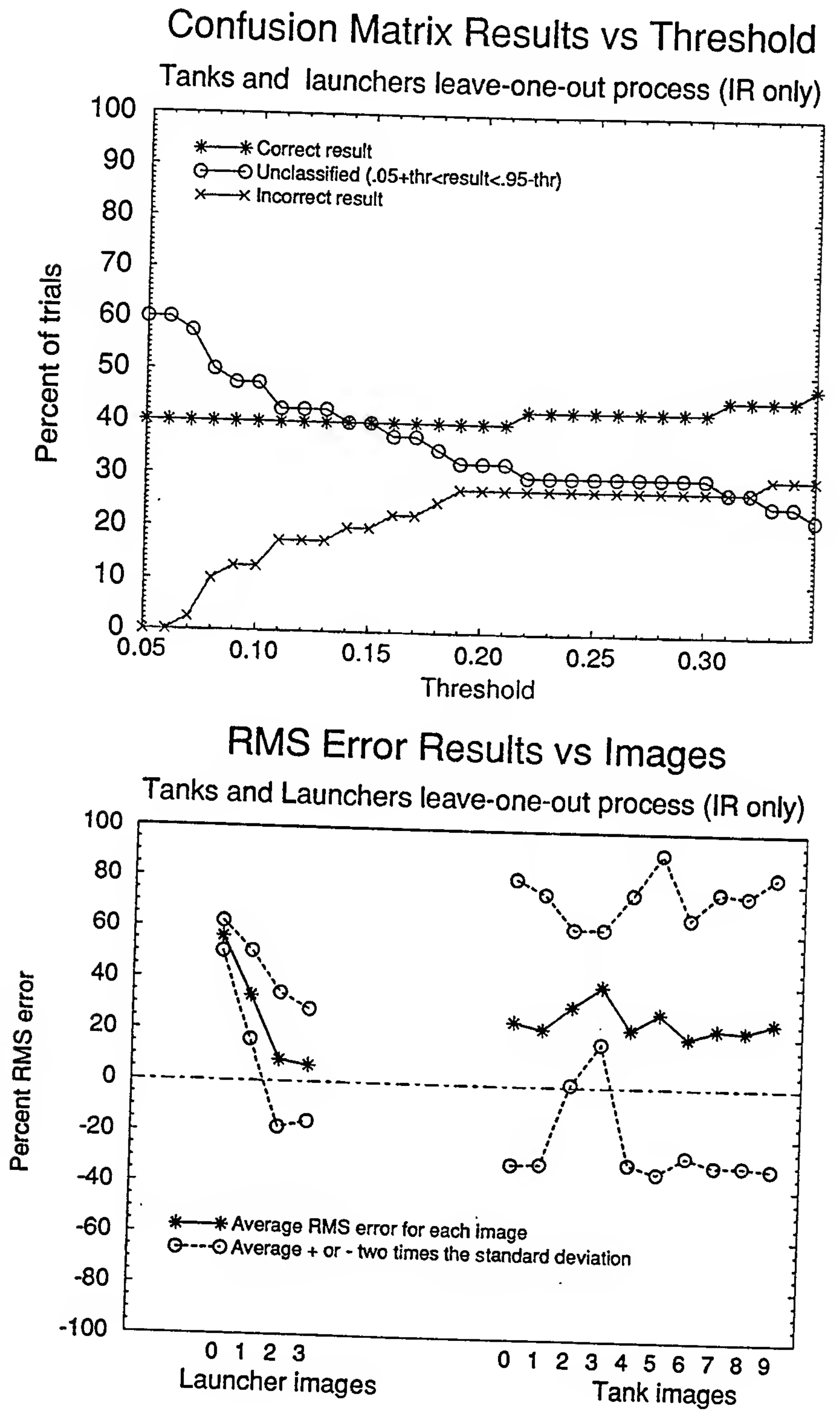
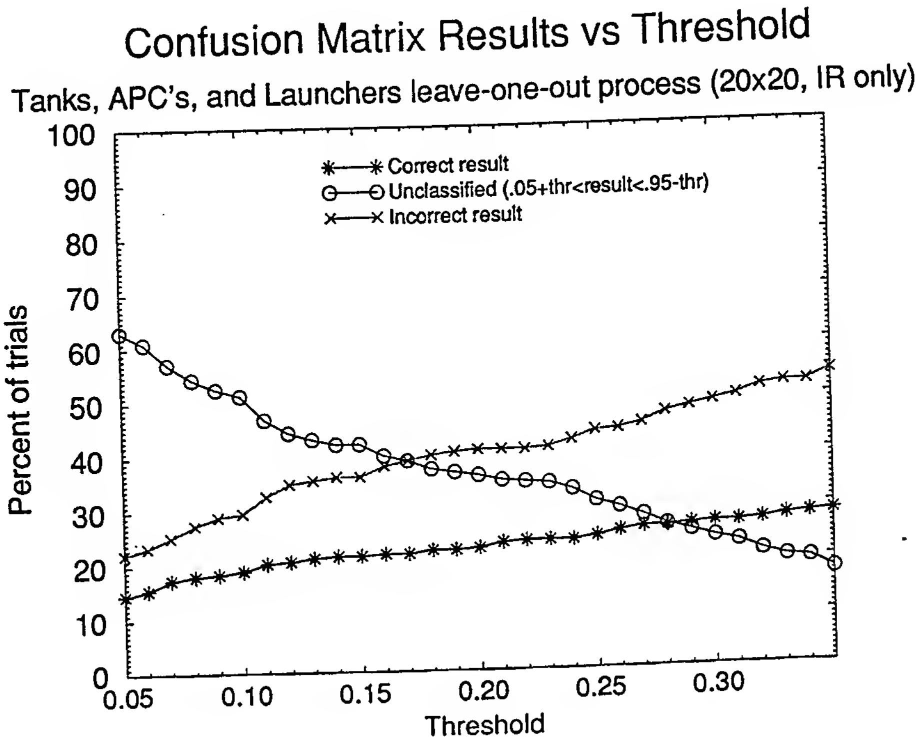


Figure 5. Neural network performance on two target types with no image clustering.



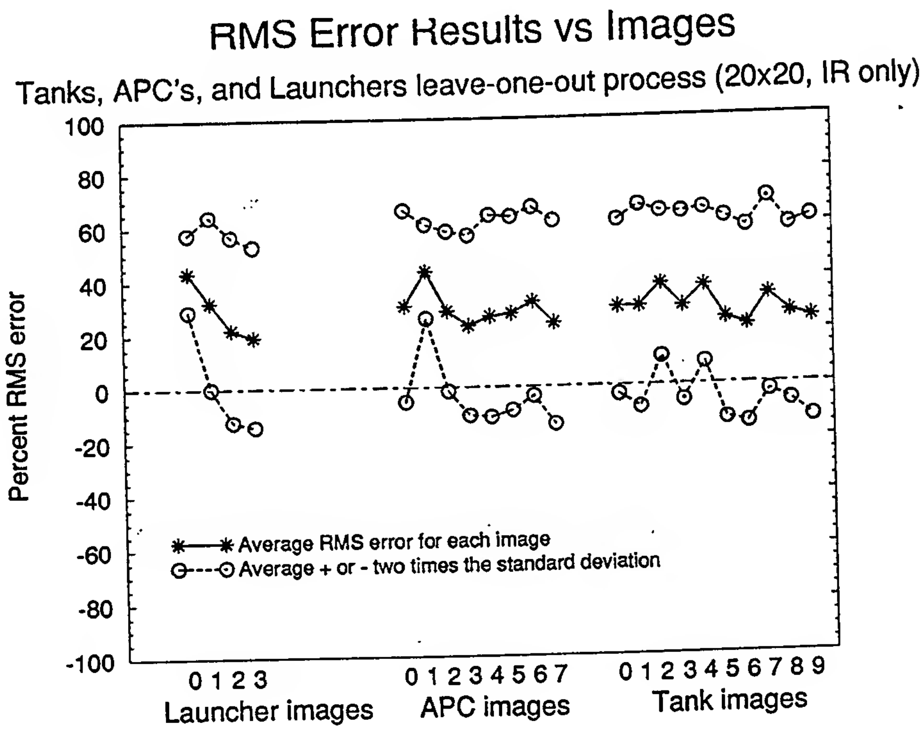


Figure 6. Neural network performance on higher resolution images. All three target types are included with no image clustering.

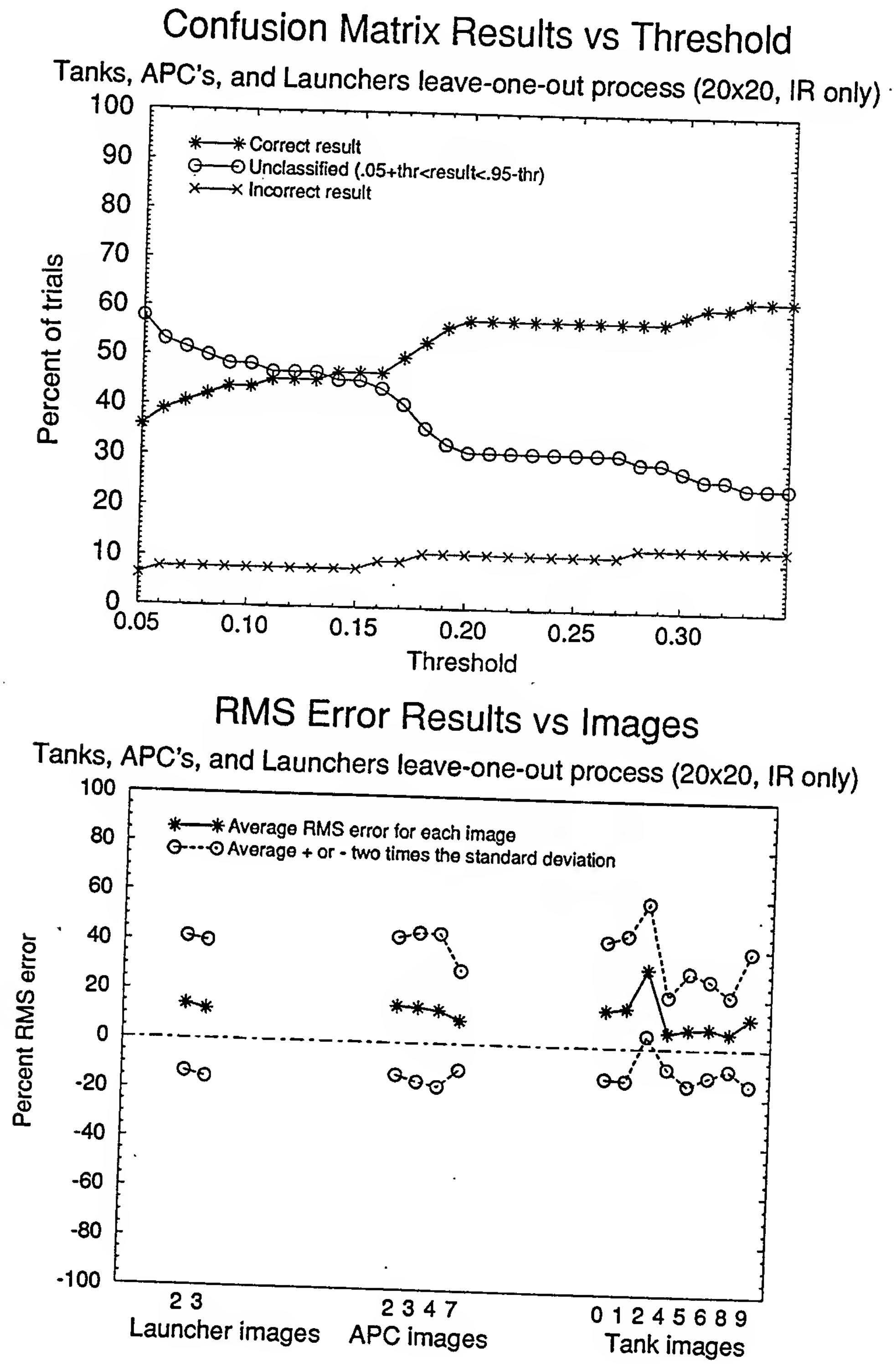


Figure 7. Neural network performance on higher resolution images. All three target types are included; image clustering is applied.

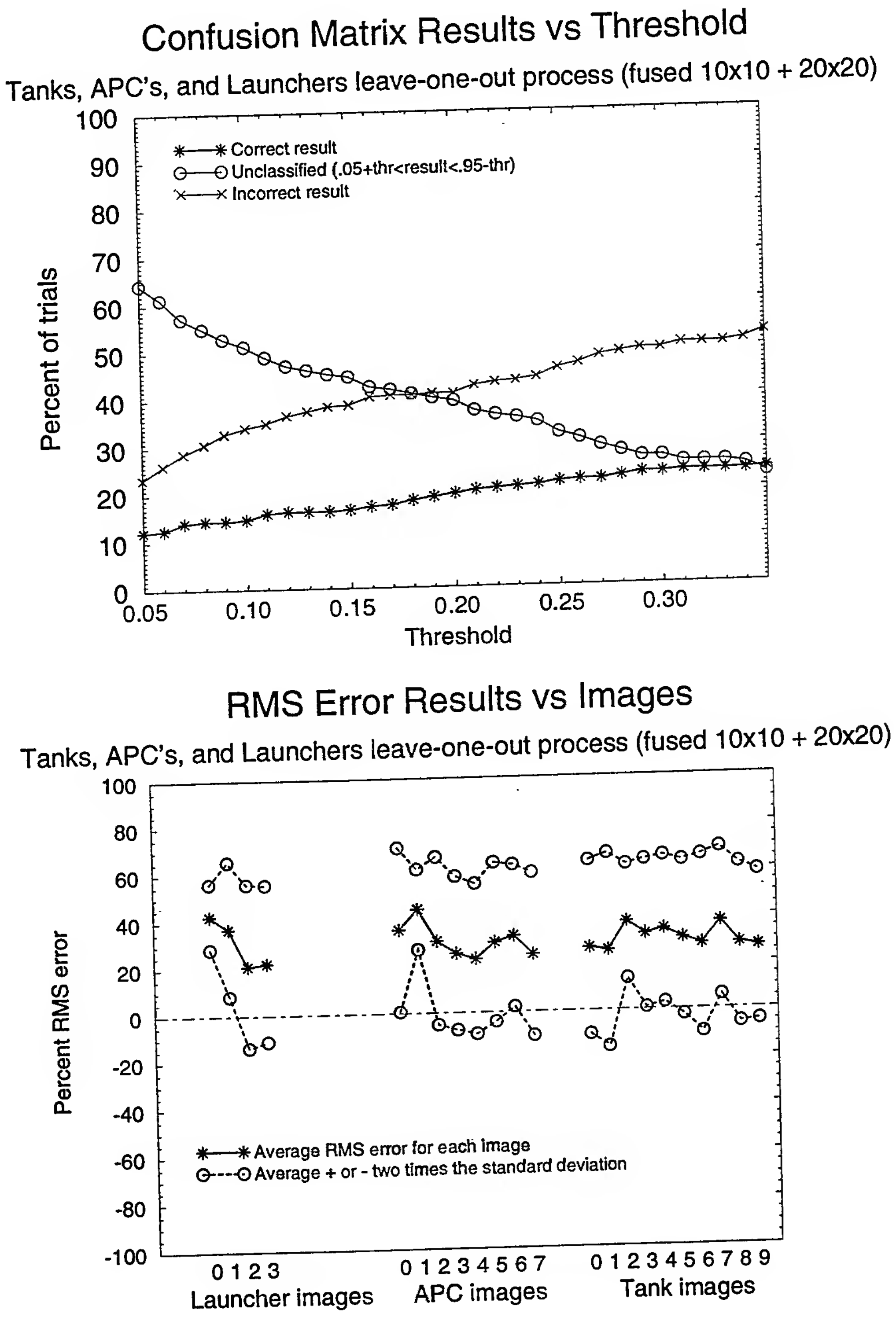


Figure 8. Neural network performance on fused low and high resolution images. All three target types are included with no image clustering.

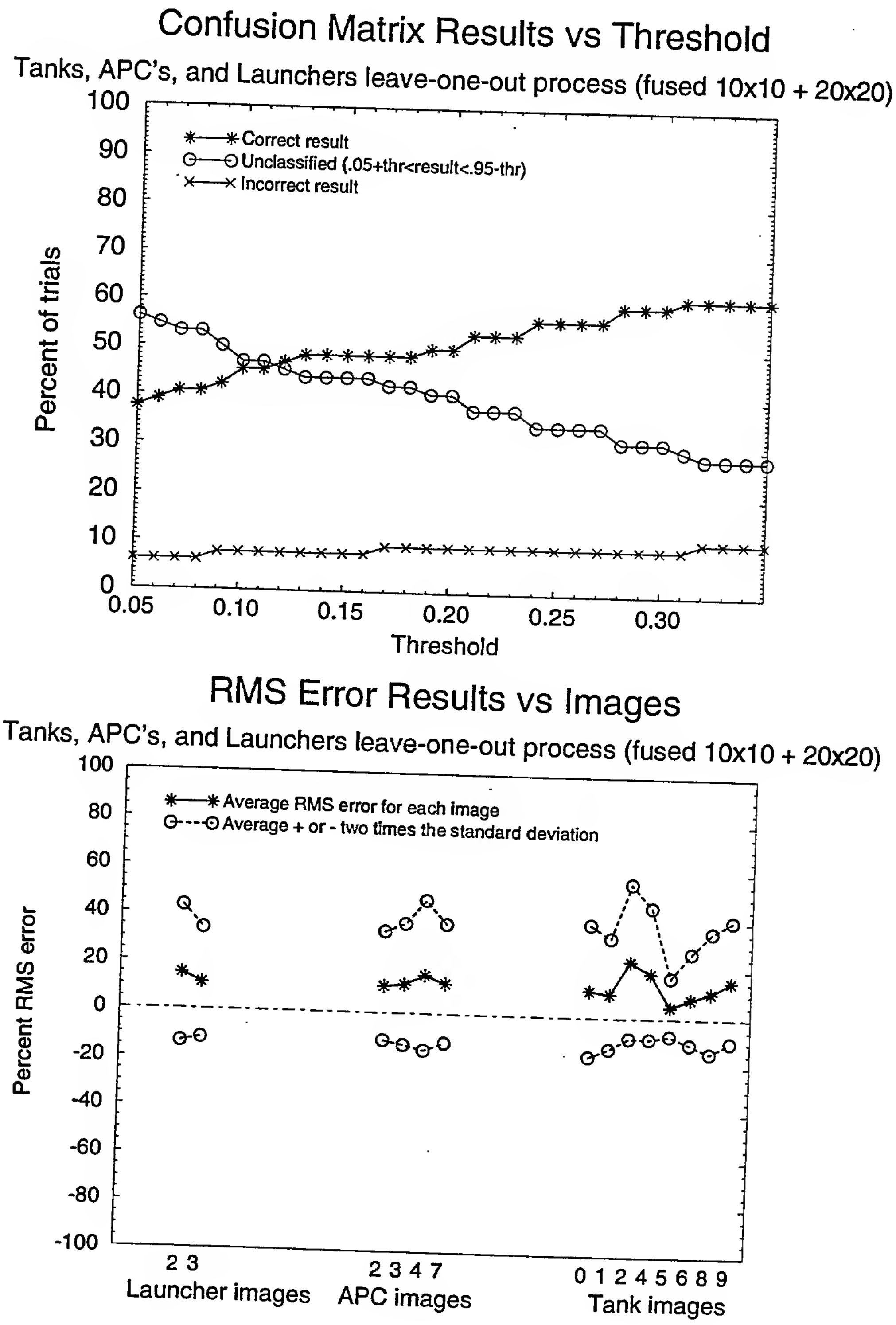


Figure 9. Neural network performance on fused low and high resolution images. All three target types are included; clustering is applied.

# CHARACTERIZING THE SOLID FRAGMENT POPULATION IN A DEBRIS CLOUD CREATED BY A HYPERVELOCITY IMPACT

William P. Schonberg
Associate Professor
Civil & Environmental Engineering Department

University of Alabama in Huntsville
Huntsville, Alabama 35899

Final Report for:

Research Initiation Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

The University of Alabama in Huntsville

April 1994

# CHARACTERIZING THE SOLID FRAGMENT POPULATION IN A DEBRIS CLOUD CREATED BY A HYPERVELOCITY IMPACT

William P. Schonberg
Associate Professor
Civil & Environmental Engineering Department
University of Alabama in Huntsville

## Abstract

The key to conducting an accurate lethality assessment is the use of a robust assessment methodology. To be applicable in a wide class of engagement scenarios, a lethality assessment methodology should incorporate all the significant response and damage mechanisms which result from hypervelocity kinetic energy weapon (KEW) impacts. One of the critical components of a lethality assessment is the characterization of the debris cloud created by the initial KEW impact. Without a proper characterization of the debris cloud material it is impossible to obtain an accurate prediction of the response of an interior target component to subsequent debris cloud impact loadings. The fast-running codes FATEPEN2, KAPP-II, and PEN4 contain fragmentation models that characterize the debris cloud fragment population resulting from a high speed impact. The objectives of the work describied in this report was first, to compare the predictions of the fragmentation models within these three codes against each other, and, second, to document similarities and differences between the predictive capabilities of the three fragmentation models over the entire 2-16 km/s impact velocity regime. These objectives were achieved through a parametric study of debris cloud material characterization using the fragmentation schemes in the FATEPEN2, KAPP-II, and PEN4 semi-empirical lethality assessment schemes over the 2-16 km/s impact velocity regime for a variety of projectile and target materials and configurations.

# CHARACTERIZING THE SOLID FRAGMENT POPULATION IN A DEBRIS CLOUD CREATED BY A HYPERVELOCITY IMPACT

William P. Schonberg

## 1.0 INTRODUCTION

The key to conducting an accurate lethality assessment is the use of a robust assessment methodology. The desire to understand the damage mechanisms that produce warhead kills in missile systems has increased the need for more accuracy in both response characterization and in lethality assessment. To be applicable in a wide class of engagement scenarios, a lethality assessment methodology should incorporate all the significant response and damage mechanisms which result from hypervelocity weapon-target interactions. For kinetic energy weapon (KEW) impacts, which can occur at speeds ranging from 2 to 16 km/s, target response and damage mechanisms can be divided into distinct categories: 'local phenomena' and 'global phenomena.'

Local response and damage phenomena are primarily due to the intense initial loading associated with a hypervelocity impact: material damage occurs very quickly (on the order of microseconds) and is limited to an area near the impact site. At sufficiently high impact velocities, shatter, melting, and/or vaporization of the materials can occur. Global response and damage phenomena occur over a longer period of time (on the order of milliseconds), under less intense loads, and over a larger area of the target structure. In KEW impacts, one or more debris clouds are created during the initial impact on the outer wall of a target. These debris clouds spread out as they move through target voids and eventually impact an inner wall or interior component of the target structure. Depending on the impact velocity and the relative material properties of the projectile and target, these debris clouds can contain solid, melted, and/or vaporized projectile and target materials.

One of the critical components of a lethality assessment is the characterization of the debris cloud created by the initial KEW impact. Without a proper characterization of the debris cloud material it is impossible to obtain an accurate prediction of the response of an interior target component to subsequent debris cloud impact loadings. Unfortunately, because very little impact test is available at speeds above approximately 8 km/s, characteristics of the debris cloud fragments created by impacts at velocities above 8 km/s have yet to be well defined. This in turn implies that semi-empirical lethality assessment schemes currently used to determine the lethal effectiveness of KEW systems are significantly limited in their characterization of the material in the debris clouds created by very high speed impacts. The need for lethality assessments of KEW impacts at speeds in excess of the maximum velocities for which currently available lethality assessment schemes are valid thus creates a dilemma for the lethality assessment community.

Most obviously, the results of lethality assessments using semi-empirical codes may be questionable at impact velocities greater than those for which the codes were designed. However, the results of hydrocode damage assessments may also be suspect in scenarios involving impact velocities greater than the highest impact velocity for which the equation-of-state used by the hydrocode is valid. Thus, the results obtained using hydrocodes may be just as correct or incorrect as those obtained using fast-running semi-empirical codes. While hydrocodes may eventually reach a level of sophistication where their equations-of-state are valid for impact velocities greater than 8 km/s, it appears that in the immediate future the most cost-effective means of performing a lethality assessment is with fast-running semi-empirical codes. The importance of performing an accurate debris cloud characterization in a lethality assessment and the availability of a variety of lethality assessment codes naturally begs the question as to whether or not the available codes are consistent in their debris cloud material characterizations. This question applies not only to the impact velocity regimes within which the codes have been experimentally verified (typically between 2 and 8 km/s), but also for speeds beyond the testable regime (i.e., in excess of 8 km/s).

The semi-empirical codes FATEPEN2 [1-4], KAPP-II [5-7], and PEN4 [8,9] contain fragmentation models that characterize the debris cloud fragment population resulting from a high speed impact. The equations in these computer codes have been designed to provide estimates for the number and sizes of the solid fragments resulting from such an impact, as well as their speeds and angular distributions about the original projectile trajectory. Each of these codes has been benchmarked with experimental test results and designated for use within a specific impact velocity regime and for specific materials and target configurations based on those tests. The objectives of the work described in this report were first, to compare the predictions of the fragmentation models within these three codes against each other, and, second, to document similarities and differences between the predictive capabilities of the three fragmentation models over the entire 2-16 km/s impact velocity regime.

A parametric study of debris cloud material characterization was performed using the fragmentation schemes in the FATEPEN2, KAPP-II (version 1.1), and PEN4.v10 semi-empirical lethality assessment schemes over the 2-16 km/s impact velocity regime for a variety of projectile and target materials and configurations. The analysis focused on the characterization of the solid fragment population in a debris cloud created by a hypervelocity impact. This included calculating the number of projectile and target material fragments, as well as their sizes, speeds, and trajectories. In addition to comparing the predictions of the various fragmentation models against one another, the three fragmentation schemes were evaluated for their completeness, their ease of implementation and use, and the availability of material properties required for their operation.

## 2.0 FRAGMENTATION SCHEMES: SYNOPSIS AND QUALITATIVE EVALUATION

The following sections contain a brief description of the three fast-running semi-empirical lethality codes and their fragmentation schemes. In each case, a brief summary of the codes and their histories is presented, followed by a summary of their respective fragmentation schemes. It is noted that no attempt is made to critique the entire lethality assessment capability of any of the three fast-running codes selected for study. The focus of this investigation and the comments made is strictly the manner in which each code treats the fragmentation of the impacting projectile and the target material in the immediate vicinity of the impact site. The comments made are also those of a first-time user of the three codes, and as such, can be especially useful to those users experimenting with these three codes for the first time.

### 2.1 FATEPEN2 -- An Introduction

The FATE family of codes [1-4] was developed for the Naval Surface Weapons Center (NSWC) for analyzing the impacts of warhead fragments against aircraft structures over an impact velocity range of 1.0 to 5.0 km/s. The present version of FATEPEN2 is actually an improvement of FATEPEN, which itself was created by combining two independently developed projectile penetration programs called FATE and PENBAM. The code FATE was developed to simulate very high velocity projectile impacts of aluminum and steel plates and focuses on the transformation of a single projectile into a debris cloud and the resulting damage to a target plate array. PENBAM was developed to simulate high velocity of multiple space plates by steel, tungsten, and DU projectiles. While PENBAM considers a variety of projectile shapes, it is concerned only with subsequent penetrations by the primary residual projectile fragment subsequent to impact. Thus, the original FATEPEN code, and, subsequently FATEPEN2, incorporates the debris cloud formation, penetration, and plate damage algorithms of FATE as well as the generality of the PENBAM program with regard to projectile shape and projectile and plate materials. The FATEPEN code has been modified over the years to include projectile tip erosion even at impact velocities below shatter velocity. The equations within the current version of FATEPEN2 predict the number of plates perforated in multi-plate target arrays as well as the diameters of the holes in the perforated plates. In addition, FATEPEN2 predicts the number, size, trajectories, and velocities of the fragments in the debris clouds created as the projectile first impacts the outermost plate and then as its remains move through the multi-plate target array.

## 2.2 FATEPEN2 -- Fragmentation Scheme

The fragmentation scheme of FATEPEN2 consists of a series of equations that calculate:

1) the ballistic limit velocity of the selected projectile/target material and geometric configuration;

2) a threshold velocity for projectile fracture, i.e., fragmentation;

3) the largest residual projectile mass in the event of fragmentation;

4) the velocity of the largest residual projectile mass;

5) two additional mean projectile fragment masses and the number of such fragments;

6) a mean target fragment mass and the number of target fragments; and,

7) a spread angle for the projectile fragments.

The expression for the ballistic limit velocity is a semi-empirical equation while the expression for the residual velocity is based on the work of Recht and Ipson [10]. The expression for projectile fracture threshold velocity is again purely empirical, and is based on NSWC test data. In calculating the largest residual projectile fragment mass, a distinction is made between impacts that result only in projectile erosion (i.e., no projectile fragmentation) and those which cause the projectile to fracture (i.e., to fragment). Unlike KAPP-II and PEN4, FATEPEN2 does not distribute the projectile fragment masses. Rather, it computes the largest residual projectile mass and average mass values for two sub-classes of projectile fragments. The equation for calculating debris cloud spread is also a semi-empirical equation and is a function of impact conditions, material properties, etc. However, the spread of the target material fragments is taken to be independent of material properties and only a function of impact angle.

The equations within the FATEPEN2 code are valid for a wide variety of metallic spherical, cylindrical, and parallelepiped projectiles impacting steel and aluminum target plates. It is stated in the FATEPEN2 documentation that the highest applicable impact velocity is 5 km/s, and that the code should not be used when the target thickness-to-projectile diameter ratio exceeds approximately 2.5. With such a limited regime of applicability, the question naturally arises as to just how well (or poorly) the code performs when it used outside its prescribed regimes of applicability.

### 2.3 Some Comments & Observations on the FATEPEN2 Fragmentation Scheme

The documentation for the FATEPEN2 code is fairly complete, but very difficult to understand. There are no sub- or super-scripts in any of the equations. This renders the equations practically unreadable and almost impossible to implement without resorting to the code itself. In addition, the equations are almost never self-contained, and rely heavily on tabular information and input from other (similarly cumbersome equations). Fortunately, the actual code was available during the course of this study so that it was possible to obtain quantitative information using the equations within FATEPEN2.

The fragmentation scheme of FATEPEN2 is not nearly as sophisticated as that of KAPP-II or PEN4. While KAPP-II and PEN4 provide an actual mass distribution for the projectile and target fragments, FATEPEN2 provides the mass of the largest residual projectile fragment and the average masses of two 'sub-classes' of projectile fragments. For the target fragments, FATEPEN2 merely presents an average fragment mass. In addition, while the spread of the projectile fragments is actually calculated, the spread of the target fragments is taken to be a function of impact angle only, regardless of material properties, impact velocity, etc.

On a more positive note, the FATEPEN2 algorithm for calculating the largest residual projectile mass is very sophisticated, including such phenomena as shock erosion, shear extrusion, and late-stage erosion. In addition, the FATEPEN2 documentation does an excellent job describing the limitations of the code with respect to materials, geometry, etc., and the code itself is applicable over a relatively wide range of projectile material/target material combinations. However, as noted previously, its range of applicability in terms of impact velocity is rather limited.

### 2.4 KAPP-II (version 1.1) -- An Introduction

KAPP-II was developed for the Defense Nuclear Agency to predict damage to complex three-dimensional targets impacted by multiple hypervelocity projectiles, including chunky fragments, rods, and hollow cylinders [5]. It is the fusion of the previously developed KAPP and KNAPP computer codes ([6] and [7], respectively). KAPP-II has been calibrated for specific projectiles against specific target types with an extensive experimental database covering an impact velocity regime of approximately 1 to 9 km/s. The code contains algorithms that have been modularized and arranged to allow a user to address a wide range of impact problems.

The algorithms within KAPP-II can be placed in one of two categories: projectile algorithms, which characterize the state of the projectile as it passes through a target; and, target algorithms, which characterize the response of target components. Projectile state characterization includes effects such as erosion and fragmentation. Target response characterization includes cratering, component dismemberment, penetration, and perforation. The semi-empirical relationships within KAPP-II allow the user to characterize the state of the projectile as it passes through the target as well as the response of the target system to the impact loading of the initial projectile and the debris created by the initial impact. In an effort to be a powerful and flexible tool, KAPP-II allows the user to select algorithms believed to be appropriate for the impact scenarios of interest.

### 2.5 KAPP-II (version 1.1) -- Fragmentation Scheme

KAPP-II consists of a series of algorithms with equations that can be used to calculate:

1) the largest residual projectile mass with and without projectile erosion;

2) the velocity of the residual projectile mass;

3) the number of projectile fragments;

4) the number of target fragments;

5) the mean projectile fragment mass;

6) the mean target fragment mass;

7) a distribution of the projectile fragment masses;

8) a distribution of the target fragment masses; and,

9) the spread angle of the debris cloud containing projectile and target fragments.

Upon review of the KAPP-II documentation, it becomes immediately apparent that there is no explicit expression for the ballistic limit of the configuration and conditions under consideration. Such a determination is made based on a comparison of the results of a penetration depth equation and the thickness of the target plate, which is a highly subjective process. Thus, when using the equations within KAPP-II to determine a residual velocity, the user must first determine whether or not perforation has occurred. For long rods, the expression for residual velocity is based on the distance from the projectile impact face at which the rarefaction wave overtakes the impact shock in the projectile. It reflects the average of the initial front end and the initial rear end velocity at impact. The expression for short rods is a semi-empirical expression.

The largest residual projectile fragment mass can be calculated in one of two ways. First, the user can assume that the entire projectile will be fragmented. While this may be a reasonable assumption for disk-like and 'chunky' projectiles, it is generally not appropriate for long rod projectiles. Nonetheless, under such an assumption, the largest residual projectile fragment is then simply the largest fragment mass in the Weibull distribution of the projectile mass fragments. Second, the user can first calculate the length of the impacting projectile that is eroded by the initial phase of the impact event. By subtracting the eroded projectile mass, the largest residual projectile mass is immediately obtained. The eroded projectile mass can then be fragmented using the standard KAPP-II approach. If the entire projectile length is eroded, then the entire projectile can assumed to be fragmented, which is, in effect, the first option just discussed.

Although there appear to be four different expressions available for calculating debris cloud spread angle, three of them are calibrated to calculated spaced penetration depth and not impact angle. The expression that actually does predict a spread angle is also based on the distance from the leading edge of the projectile at which the rarefaction wave overtakes the impact-induced shock wave in the projectile.

## 2.6 Some Comments & Observations on the KAPP-II Fragmentation Scheme

The equations within the KAPP-II fragmentation scheme are easy to understand, implement, and use, fairly self-contained in that their reliance on tabular data is minimal, and applicable over a wide variety of target and projectile materials, projectile shapes (long and short rods, solid and hollow rods, spheres, etc.), and target configurations (flat plates, re-entry vehicles, etc.). The user is presented with a wide choice of algorithms to use and is also provided with information regarding appropriate algorithm groupings. This is a very useful piece of information for the first-time user.

One of the most serious drawbacks of the equations within KAPP-II is the lack of any information regarding their limitations. While some general statements are possible, specific statements concerning individual equations are not possible because not all of the equations were calibrated over the same dataset. In addition, while equations are provided for 'long' and 'short' rods, the user is left to wonder as to the value of the length-to-diameter ratio that corresponds to the long rod/short rod boundary. Unfortunately, no guidance is offered in the KAPP-II documentation as to this matter.

While the applicability of the equations within KAPP-II extends to a wide variety of metallic and non-metallic materials, the use of the equation that calculates the number of fragments is complicated by the fact that it requires as input the 'ultimate failure strain' of the material. This is a non-standard material property and is not readily found in standard materials properties' handbook. To make matters worse, no definition of this property is provided in the documentation regarding its origin or definition. Hence, while this property is given for the materials in the properties table provided at the end of the KAPP-II documentation, adding additional materials to this table for use with the fragmentation equation is impossible. In addition, the manner in which the 'ultimate failure strain' parameter is used in the equation to calculate the number of fragments created is also misleading. The user must actually divide the parameter value by 100 prior to its substitution into the equation. No where is this stated in the KAPP-II documentation: this fact becomes apparent only after consulting the KAPP-II source code, which fortunately was available for this investigation.

2.7 PEN4.v10 -- An Introduction

The PEN4 lethality assessment code was developed for the NSWC in attempt to model the impact of high density metal projectiles against multi-plate target arrays consisting of thin aluminum plates over a wider range of impact velocities [9]. It is a collection of equations describing the projectile state and target impact response. PEN4 is a modular code, constructed around the various physical phenomena that occur during a high speed impact event. Thus, any improvement in the understanding of a particular aspect of the impact process requires replacing only one module and does not require any extensive rewriting of the entire code. This model is similar to the FATEPEN2 model in that the equations within PEN4 were derived using a number of simplifying assumptions and experimentally derived constants.

By restricting the lower limit of the impact velocity to approximately 3.5 km/s, PEN4 neglects shear failure in the projectile material; by restricting its upper limit of applicability to 7.6 km/s, PEN4 neglects material melting and vaporization. In the most recent version [9], PEN4 has been updated to include more advanced fragmentation schemes (see, e.g., [11]). These fragmentation models are a considerable improvement over the models used in the earlier versions of the code (see, e.g., [8]). Additional recent improvements in PEN4 include new relations for hole diameter, crater depth and diameter, and largest residual projectile fragment mass.

2.8 PEN4.v10 -- Fragmentation Scheme

The fragmentation scheme of PEN4 consists of a series of equations that calculate:

1) the ballistic limit velocity of the selected projectile/target material and geometric configuration;

2) a threshold velocity for projectile fragmentation;

3) the largest residual projectile mass;

4) the velocity of the largest residual projectile mass;

5) the mean mass of the projectile fragments;

6) a distribution of the projectile fragments; and,

7) the spread of the debris cloud containing the projectile fragments.

The expression for the ballistic limit velocity is based on the form of the expression developed in Project THOR [12,13]. The development of the expression for the residual velocity is a generalization of the original derivation of residual velocity in the event of a plugging impact event given by Recht and Ipson [10]. The expressions for fragmentation threshold velocities are purely empirical, and are based on IRAD test data. In calculating the largest residual projectile fragment mass, a distinction is made between power-law fragmentation processes and power-law roll-off fragmentation processes. This distinction is based on the observation that there is a rapid roll-off in largest residual fragment size for impact velocities above approximately 4 km/s for steel and aluminum projectiles impacting aluminum target plates. As in KAPP-II, PEN4 distributes the projectile fragment masses using a Weibull distribution function.

The equations within the PEN4 code are strictly valid for spherical aluminum 2024-T3 and 1018 steel projectiles impacting aluminum 2024-T3 target plates. It is stated in the documentation that projectile masses in excess of 30 grams and values of target thickness-to-projectile diameter rations greater than 2.0 will likely yield spurious results. Thus, the equations within PEN4 are applicable only to a limited class of problems. However, it would still be instructive to ascertain how well (or poorly) they perform when used outside their specified regimes of applicability.

### 2.9 Some Comments & Observations on the PEN4 Fragmentation Scheme

The current documentation available for the PEN4 code is easy to follow and understand. The equations are fairly easy to implement and use; the usual preventative care must be taken to ensure that appropriate units are used. Unfortunately, the current documentation does not provide an example using the equations to guide the first-time user (unlike an early report [8] which did). Two minor errors were found in the debris cloud spread equation and in the equation governing the boundary between power-law and power-law roll-off fragmentation. Similar to KAPP-II but unlike FATEPEN2, PEN4 actually provides a mass distribution for the projectile fragments created as a result of the impact event.

The limitations of the code are fairly well documented with regard to acceptable projectile and target materials and configurations (i.e., target thickness, projectile diameter, etc.). In fact, the comments regarding the validity of the equation for mean projectile fragment mass are quite frank (it is stated outright that the results it provides are dubious at best). However, while most likely an error of omission, the latest documentation of PEN4 does not provide an expression for calculating the total number of projectile and target fragments. In addition, no information is provided to explicitly address the fragmentation of the material ejected from the target plate. Thus, because the actual PEN4 code was not available for this investigation, it was later not possible to compare these two predictive aspects of the PEN4 fragmentation scheme against the predictions of the corresponding algorithms within KAPP-II and FATEPEN2.

Finally, the equation for residual velocity is based on a plugging event; its usefulness in predicting debris cloud motion is therefore rather dubious. While various forms of the equation are given in the documentation (e.g.: projectile and target plug travel together or travel separately; the projectile is sharp or blunt; the projectile is deformable or rigid), no guidance is provided to the user as to which options are likely to be most valid in a particular impact scenario of interest. On a more positive note, the expression for residual velocity does contain the ballistic limit velocity, as well as impact velocity, as input variables. If the impact velocity is less than the ballistic limit velocity, the residual velocity will be an imaginary number, a clear indicator to the user that there is no need to proceed any further in the calculations. As noted previously, this characteristic is shared by the corresponding equation in FATEPEN2, but not by KAPP-II.

## 3.0 FRAGMENTATION SCHEMES: QUANTITATIVE COMPARISONS

### 3.1 Introductory Comments

The predictions of the fragmentation schemes of the three semi-empirical codes were compared with each other by obtaining projectile and target material fragment characteristics for impact velocities ranging from 1 to 15 km/s with right circular cylinder projectiles normally impacting flat target plates without any projectile yaw. For each impact velocity, three projectile diameter-to-target plate thicknesses (D/T) ratios were considered: D/T=0.1, D/T=1.0, and D/T=10.0; and, for each D/T ratio, three projectile length-to-diameter ratios were considered: L/D=0.1, L/D=1.0, L/D=10.0. In this manner, the codes were used to generate information for disk, compact, and long rod projectiles whose diameters were much smaller than, the same size as, and much larger than the thickness of the target plate. In order to be able to use the PEN4 equations, projectile materials considered were a high strength steel and a generic aluminum while the target material considered was a generic aluminum.

The following general observations were made after examining the outputs obtained from FATEPEN2 and the program written to implement the PEN4 equations:

1) Both PEN4 and FATEPEN2 agreed that none of the aluminum projectiles would perforate the aluminum target plate when D/T=0.1; both also agreed that the steel projectile with L/D=10.0 would perforate the aluminum target plate when D/T=0.1, but only at impact velocities of 5 km/s and above. However, unlike FATEPEN2, PEN4 also predicted that the steel projectile with L/D=1.0 would perforated the target plate when D/T=0.1, but only at speeds above 7 km/s.

2) While both FATEPEN2 and PEN4 predicted that both steel and aluminum projectiles with L/D=10.0 would perforate the aluminum target plate when D/T=1.0, there were numerous differences between the predictions of FATEPEN2 and PEN4 with regard to whether or not both the steel and the aluminum projectiles with L/D=0.1 and L/D=1.0 would perforate the aluminum target plates when D/T=1.0.

3) Both FATEPEN2 and PEN4 agreed that all three types of the steel and aluminum projectiles would perforate the aluminum target plates when D/T=10.0

Based on this information, it would appear that with regard to perforation or no perforation of the target plates, there was more agreement than disagreement between PEN4 and FATEPEN2 for the materials considered.

### 3.2 Largest Residual Projectile Fragment Mass

Figures 1 through 4 show plots of non-dimensional largest residual projectile mass as a function of impact velocity for the al-on-al impacts (Figures 1 and 2) and the st-on-al impacts (Figures 3 and 4). A cursory examination of these four figures that, with the exception of KAPP-II (with erosion) when L/D=10.0, all three codes predict very small maximum residual projectile fragments above impact velocities of 3 km/s for al-on-al impacts and 5 km/s for st-on-al impacts. For PEN4, this is no doubt due to the fact that the power-law/power-law roll-off boundaries are approximately 3 km/s for al-on-al impacts and approximately 4.5 km/s for st-on-al impacts. It is also noted that the maximum residual projectile mass as predicted by FATEPEN2 when L/D=10.0 decreases at a slower rate than the other projectile/target combinations, but eventually does drop to a very small value.

The predictions of KAPP-II are interesting in that if projectile erosion is considered when L/D=10.0 and D/T=10.0, then KAPP-II predicts that the (relatively massive) projectile passes through the (relatively thin) target plate relatively unscathed. This also appears to be true when L/D=10.0 and D/T=1.0, although to a somewhat lesser degree. However, if projectile erosion is ignored, then the predictions of KAPP-II are of the same order as those of PEN4. Unfortunately, knowing the amount of erosion *a priori* is impossible, and the user is given no guidance by the KAPP-II documentation as to when to consider projectile erosion and when to ignore it. Common sense would appear to dictate that when a long rod (i.e., L/D=10.0) impacts a relatively thin (i.e., D/T=10.0) target plate, it should pass through relatively unscathed, especially at very high impact velocities (i.e., in excess of 10 km/s).

Closer examination of the data generated by the three codes does reveal some more significant discrepancies in the predictions of largest residual projectile mass. The most serious of these is that while the smallest maximum residual fragment masses predicted by KAPP-II are on the order of 0.1% to 2% of the initial projectile mass, FATEPEN2 and PEN4 predict maximum residual projectile masses for similar impact conditions that are on the order of $10^{-5}$ times the original projectile mass. Thus, FATEPEN2 and PEN4 in effect predict a more 'thorough' or 'complete' fragmentation of the projectile than does KAPP-II.

### 3.3 Projectile Residual Velocity

Figures 5 through 8 and 9 through 14 show plots of non-dimensional residual velocity for al-on-al and st-on-al impacts, respectively. In calculating residual velocity using KAPP-II, only the short rod and long rod equations were used for L/D=0.1 and L/D=10.0 projectiles, respectively. However, the predictions of both the long rod and the short rod equations in KAPP-II were obtained and plotted for L/D=1.0 projectiles (see, e.g., Figure 6). It is anticipated that the actual residual velocity when L/D=1.0 according to KAPP-II would lie somewhere between the predictions of the short rod and long rod equations. In addition, data from two extreme cases were obtained and plotted using PEN4: elastic impact (e=1) and inelastic impact (e=0). Again, it is anticipated that the actual residual velocity according to PEN4 would lie somewhere between the two values.

As can be seen in Figures 5 and 6 (i.e., al-on-al when D/T=1.0), the predictions of KAPP-II and FATEPEN2 when L/D=1.0 and 10.0 are within approximately 20% to 25% of each other while the predicted values of residual velocity according to PEN4 when L/D=1.0 and 10.0 are approximately 50% of those predicted by KAPP-II and FATEPEN2. When D/T=1.0 and L/D=0.1, the predictions of KAPP-II are of the same order as those of PEN4 for al-on-al impacts. In Figures 7 and 8 (i.e., al-on-al when D/T=10.0) the predictions of residual velocity are significantly more uniform for the three L/D ratios considered, but especially for L/D=10.0 (Figure 7, top set of curves). In addition, the values predicted by PEN4 when e=0 and when e=1 differ significantly except when D/T=10.0 and L/D=10.0. Analogously, the predictions of the KAPP-II long rod and short rod equations for L/D=1.0 are nearly identical when D/T=10.0.

Figures 9 through 14 (i.e., st-on-al impacts) show the same trends in the data as those which were observed in the al-on-al impacts: fairly close agreement between KAPP-II and FATEPEN2 when L/D=1.0 and 10.0 for D/T=1.0 (and 0.1 as well); a wide gap between the predictions of PEN4 and those of FATEPEN2 and KAPP-II when L/D=1.0 and 10.0 for D/T=1.0 (and 0.1 as well); and, general agreement among all three codes when D/T=10.0, and especially when L/D=10.0. Additionally, significant differences between long and short rod predictions were found for L/D=1.0; the differences were relatively minor when D/T=1.0, and all but disappeared when D/T=10.0. Finally, in a manner again similar to the al-on-al impacts, the differences between the predictions of the e=0 and e=1 options in PEN4 were fairly significant except when L/D=10.0 and L/D=10.0

### 3.4 Materials Fragmentation

Figures 15,16 and 17,18 show plots of the number of the number of projectile and target fragments generated, respectively, by al-on-al impacts. Several features evident in these four plots highlight some of the key differences in the way KAPP-II and FATEPEN2 fragment the projectile and target materials. As noted previously, the available PEN4 documentation does not provide for a means of calculating the number of fragments generated; hence, no comparisons with PEN4 were possible.

One characteristic of the KAPP-II predictions is immediately apparent: the equation for number of fragments generated does not distinguish between projectile and target material for like-on-like impacts.

However, it should be noted that the individual fragment masses created will in fact be different for the projectile and target materials because of the difference between the initial projectile mass and the mass of the target material ejected by the impact. Another feature immediately apparent from the four sets of curves is that the number of fragments generated according to the KAPP-II equation is independent of the value of the D/T and L/D ratios; FATEPEN2, however, does allow for a distinction in geometry in calculating the number of fragments generated. Naturally, even though the number of fragments may be the same according to KAPP-II for two different values of L/D or D/T, the actual masses of the fragments formed in the two cases will still be different due to differences in to total initial mass values.

With the exception of Figure 16, in which the thickness of the target plate is much smaller than the diameter of the impacting projectile, KAPP-II predicts that a far greater number of projectile and target fragments will be created as compared to the predictions of FATEPEN2. In Figure 16, the FATEPEN2 plots are terminated at impact velocities of 7 km/s and 9 km/s for L/D ratios of 1.0 and 10.0, respectively. This is not to say that projectile fragmentation no longer occurred at impact speeds above 7 and 9 km/s. Rather, at speeds in excess of 7 and 9 km/s, FATEPEN2 predicted that the number of projectile fragments would be on the order of $10^6$ for L/D=1.0 and $10^{16}$ for L/D=10.0. While these are astronomically high numbers, perhaps some credence may be lent to them if one considers that, with the exception of the single largest residual projectile fragment, the other projectile fragments are likely to be melted at impact velocities above 7 km/s and possibly even vaporized at speed in excess of 10 km/s. In light of this consideration, the high number of 'fragments' (droplets?) predicted by FATEPEN2 no longer seems unreasonable.

Figures 19-21 and 22-24 show plots of the number of the number of projectile and target fragments generated, respectively, by st-on-al impacts. These figures show some of the same trends in the data as those which were observed in the al-on-al impacts: KAPP-II predicts a far larger number of projectile and target fragments than does FATEPEN2 (except when D/T=10.0), and that when D/T=10.0, the number of fragments predicted by FATEPEN2 become astronomically high. A major difference between these plots and the al-on-al plots is that when the projectile and target materials are not the same, KAPP-II will in fact predict a different number of fragments created for the projectile and for the target materials.

### 3.5 Debris Cloud Spread Angle

Figures 25 and 26 show plots of the debris cloud half angles for the al-on-al impacts. For the two D/T values shown, the predictions of the three codes are generally within 10° to 15° of each other. The predictions of PEN4 approach their asymptotic limit of 26° at impact speed of approximately 9 and 15 km/s for D/T=1.0 and 10.0, respectively; in both cases, the predictions of FATEPEN2 appear to approach an asymptotic limit of approximately 32° at an impact speed of approximately 7 km/s. In complete contrast to the asymptotic behavior exhibited by both the FATEPEN2 and the PEN4 predictions, the predictions of KAPP-II appear to increase without bound for D/T=1.0 and actually begin to decrease slightly for D/T=10.0 beyond a speed of approximately 12 km/s.

While an asymptotic value of debris cloud spread may be desirable in a lethality assessment code from the standpoint of predicting a maximum damage area in subsequent target components, the behavior exhibited by the KAPP-II predictor equation is probably more realistic. The reason for this, especially in the case of very thin target plates (e.g., D/T=10.0), is as follows. As the impact velocity is increased beyond the incipient fragmentation velocity of the projectile, the debris created will naturally spread out more and more. However, at very high impact speeds, it is logical to presume that a projectile impacting a very thin target plate will pass through the plate relatively unscathed, which in effect constitutes a minimal spread in the debris created. Hence, if a function is to describe the spread of the fragmented material in terms of impact speed, it must be monotonically increasing from a value near zero for low impact speeds and it must be near zero for very high impact speeds. The only way for this to happen is if it was to peak at some impact velocity in between. Naturally, the velocity at which the peak occurs, as well as the velocity at which the debris cloud spread finally settles on a near-zero value, is a function of projectile and target material properties and the geometric parameters of the system under consideration.

Figures 27 through 29 show plots of the debris cloud half angles for the st-on-al impacts. For the three D/T values considered, the predictions of the three codes in this case are generally more widespread than in the case of aluminum projectiles. The predictions of PEN4 approach their asymptotic limit of 26° at impact speed of approximately 11, 13, and 15 km/s for D/T=0.1, 1.0, and 10.0, respectively; in both cases, the predictions of FATEPEN2 appear to approach an asymptotic limit of approximately 15° at an impact speed of approximately 7 km/s. Once again, in contrast to the asymptotic behavior exhibited by both the FATEPEN2 and the PEN4 predictions, the predictions of KAPP-II appear to increase without bound for D/T=0.1 and 1.0. For D/T=10.0, there is a marked decrease in debris could spread beyond an impact velocity of approximately 11 km/s.

## 4.0 SUMMARY

### 4.1 General Comments

A parametric study of debris cloud material characterization was performed using the fragmentation schemes in the FATEPEN2, KAPP-II (version 1.1), and PEN4.v10 semi-empirical lethality assessment schemes over the 2-16 km/s impact velocity regime for a variety of projectile and target materials and configurations. The analysis focused on the characterization of the solid fragment population in a debris cloud created by a hypervelocity impact. This included calculating the number of projectile and target material fragments, as well as their sizes, speeds, and trajectories. In addition to comparing the predictions of the various fragmentation models were compared with one another, the three fragmentation schemes were evaluated for their completeness, their ease of implementation and use, and the availability of material properties required for their operation.

### 4.2 Qualitative Analysis Summary

The following is a summary of the results of the qualitative analyses performed as part of this investigation:

1) The FATEPEN2 documentation is difficult to understand; the fragmentation equations are almost impossible to implement and use without the actual source code. The PEN4 and KAPP-II documentations are easy to follow and the equations are relatively easy to implement, although some errors and omissions were observed in the PEN4 documentation.

2) The FATEPEN2 and PEN4 documentations are diligent about stating the limitations of the respective codes and the equations contained therein; the KAPP-II documentation is often negligent in this obligation to its users.

3) Both KAPP-II and PEN4 provide actual distributions of the fragment masses created while

FATEPEN2 does not. While this is not a critical flaw on the part of FATEPEN2, the other two codes do present a more complete picture of the fragmentation process.

4) While the equations within FATEPEN2 take into account the effects of the D/T and L/D ratios in determining the number of resulting fragments, those in KAPP-II do not. Ultimately, a few well-instrumented tests will be needed to decide whether or not this a problem with KAPP-II or just a bonus provided by FATEPEN2. In addition, one of the material parameters required by KAPP-II is a non-standard property. Hence, applying the KAPP-II fragmentation equation to materials not presently in its material library is impossible.

5) The FATEPEN2 algorithm for calculating the largest residual projectile mass is very sophisticated, while the KAPP-II algorithm is rudimentary in nature. Since the corresponding PEN4 algorithm is purely empirical, it may be argued that it is simplistic.

6) While both KAPP-II and FATEPEN2 provide sample problems as part of their documentation, PEN4 does not. This makes it difficult for first-time users of PEN4 to know whether or not they are interpreting the subtletics of the code properly and using it correctly.

7) No guidance is given in the KAPP-II and PEN4 documentations regarding some of the subtleties in their equations, such as whether to use the long rod or short rod equations in KAPP-II when L/D=1.0 and when to assume an elastic impact and when to assume a plastic impact in PEN4. This makes it difficult to use to equations in which these subtleties occur as predictive tools without several iterative runs.

4.3 <u>Quantitative Analysis Summary</u>

The following is a summary of the results of the quantitative analyses performed as part of this investigation:

1) <u>Perforation Resistance</u> -- There was general agreement between PEN4 and FATEPEN2 with regard to whether or not the aluminum target plate was perforated by the impacting steel and aluminum projectiles.

2) <u>Largest Residual Projectile Mass</u> -- With the exception of KAPP-II when L/D=10.0 for D/T=10.0 and when projectile erosion was included, there was general agreement between the three codes in predicting the mass of the largest residual projectile fragment relative to the mass of the original

projectile. Except as noted, all three codes predicted very small maximum residual fragments for impact velocities greater than approximately 5 km/s. For the conditions noted, the KAPP-II equations predicted that the projectile would pass through the target plate relatively unscathed.

3) Residual Projectile Velocity -- KAPP-II and FATEPEN2 agreed within 20%-25% for compact and long rod projectiles impacting moderately thick plates; all three codes were in close agreement (i.e., less than 10%) for very thin target plates.

4) Materials Fragmentation -- Except for the impact of very thin plates, KAPP-II predicted that significantly more fragments would be generated than did FATEPEN2 (i.e., by several orders of magnitude). For very thin plates, the opposite occurred: FATEPEN2 predicted that a tremendously high number of fragments would be generated. The numbers predicted by FATEPEN2 in such cases exceeded those predicted by KAPP-II by several tens of orders of magnitude.

5) Debris Cloud Spread -- All three codes agreed in their prediction of debris cloud spread within $10^{o}$ to $15^{o}$ for the al-on-al impacts. However, the differences grew to $20^{o}$ to $25^{o}$ for the st-on-al impacts.

## 5.0 REFERENCES

1. Yatteau, J.D., High Velocity Multiple Plate Penetration Model, NSWC-TR- 82-123, Dahlgren, Virginia, February, 1982.

2. Yatteau, J.D., Modifications to Program FATE - Fragment Residual Mass Calculations, Final Report, Denver Research Institute, University of Denver, Denver, Colorado, May, 1983.

3. Yatteau, J.D., Zernow, R.H., and Recht, R.F., Compact Fragment Multiple Plate Penetration Model, Volume I: Model Description, NSWC-TR-91-399, Dahlgren, Virginia, January, 1991.

4. Yatteau, J.D., Zernow, R.H., and Recht, R.F., Compact Fragment Multiple Plate Penetration Model, Volume II: Computer Code User's Manual, NSWC-TR-91- 399, Dahlgren, Virginia, January, 1991.

5. Greer, R., and Hatz, M., KAPP-II User's Manual, Version 1.1, Kaman Sciences Corporation, K92-17U(R), Colorado Springs, Colorado, April, 1992.

6. Snow, P., KAPP - Kaman Analytical Penetration Program, Kaman Sciences Corporation, K85-7U(R), Colorado Springs, Colorado, 1985.

7. Cohen, L., Kaman New Analytical Penetration Program (KNAPP) Space-Based Interceptor Modelling Effort, AFATL-TR-90-02, Eglin AFB, Florida, February, 1990.

8. Henderson, B.J., and Zimmerschied, A.B., Very High Velocity Penetration Model, NSWC-TR-83-189, Dahlgren, Virginia, May, 1983.

9. Bjorkman, M.D., Geiger, J.D., and Wilhelm, E.E., Space Station Integrated Wall Design and Penetration Damage Control, Task 3: Theoretical Analysis of Penetration Mechanics, Boeing Aerospace Corporation, Final Report, Contract NAS8-36426, Seattle, Washington, July, 1987.

10. Recht, R.F. and Ipson, T.W., "Ballistic Perforation Dynamics", Journal of Applied Mechanics, Vol. 30, No. 3, pp. 284-290, 1963.

11. Grady, D.E., "Local Inertial Effects in Dynamic Fragmentation", Journal of Applied Physics, Vol. 53, No. 1, pp. 322-325, 1982.

12. THOR 41: A Comparison of the Perforation of Fragments of Four Materials Impacting Various Plates, Ballistics Analysis Laboratory, The Johns Hopkins University, Baltimore, Maryland, May, 1959.

13. THOR 47: The Resistance of Various Metallic Materials to Perforation by Steel Fragments; Empirical Relations for Fragment Residual Velocity and Residual Weight, Ballistics Analysis Laboratory, The Johns Hopkins University, Baltimore, Maryland, May, 1959.

FIGURE 1

RESIDUAL PROJECTILE MASS

AL-->AL, D/T=1.0, L/D=1.0 & 10.0

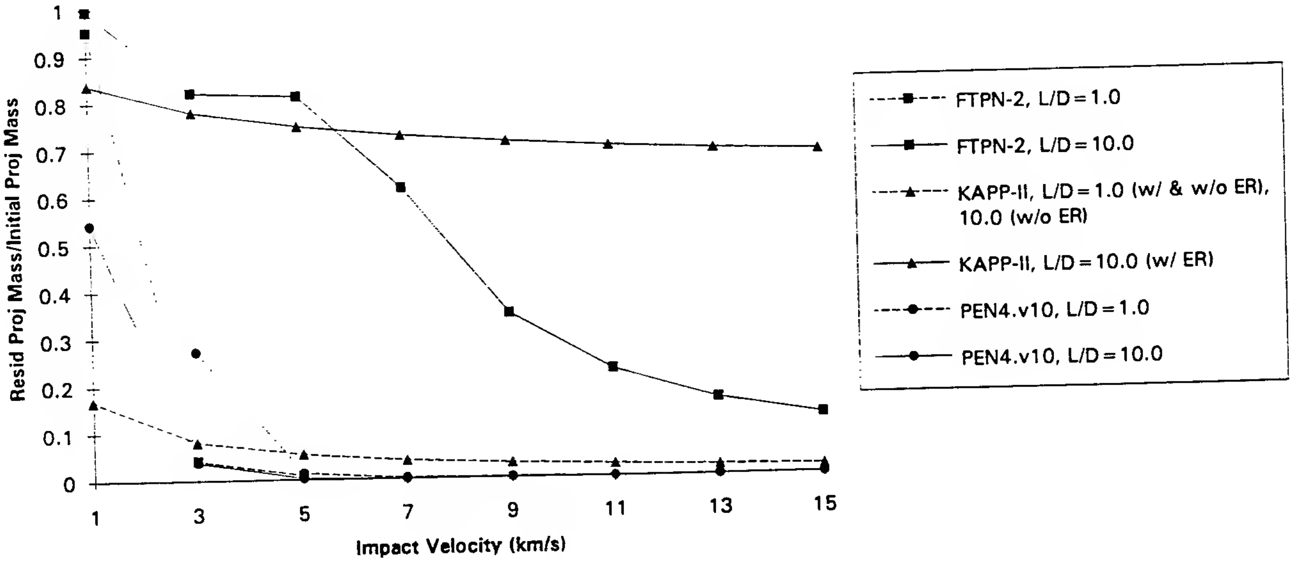


FIGURE 2

RESIDUAL PROJECTILE MASS

AL-->AL, D/T=10.0

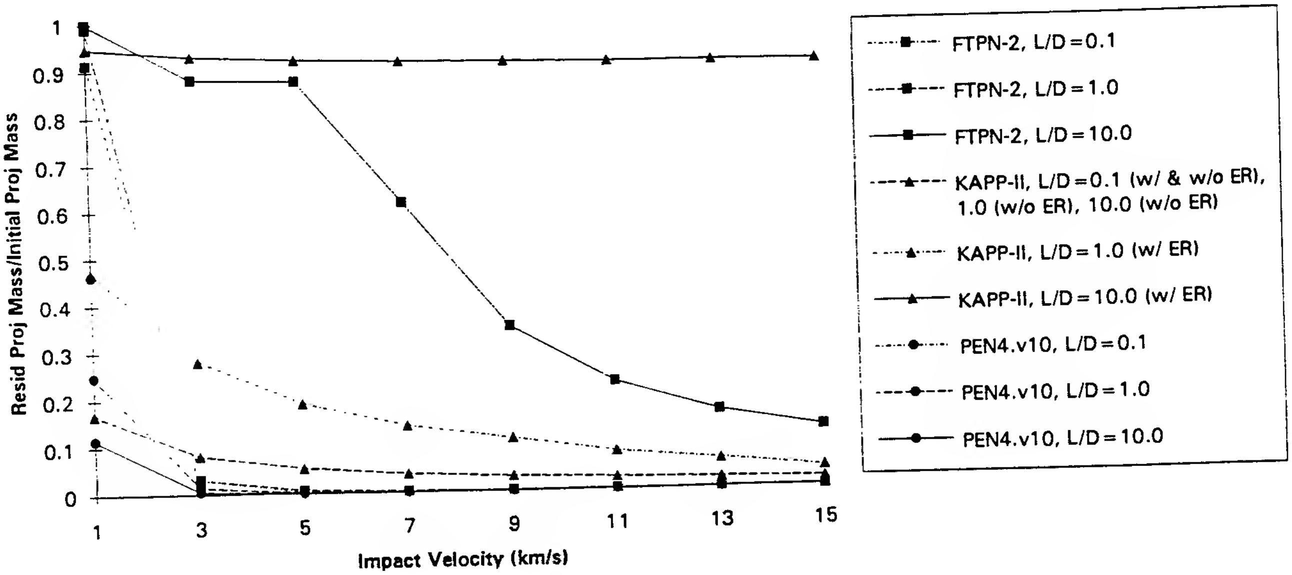


FIGURE 3
RESIDUAL PROJECTILE MASS
ST-->AL, D/T=1.0

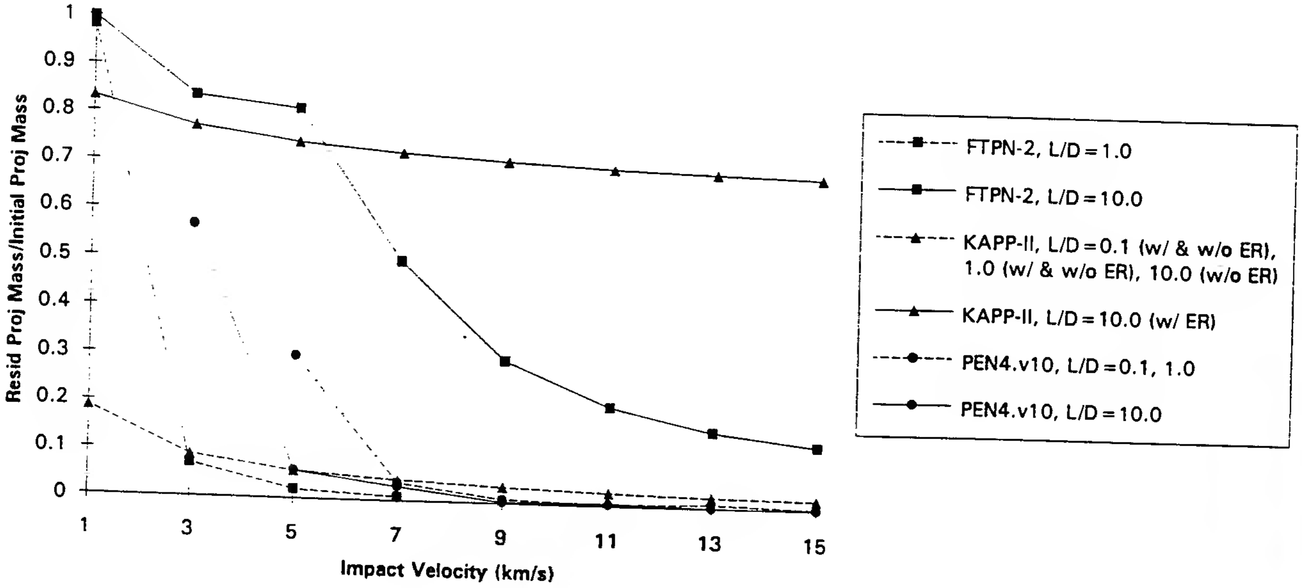
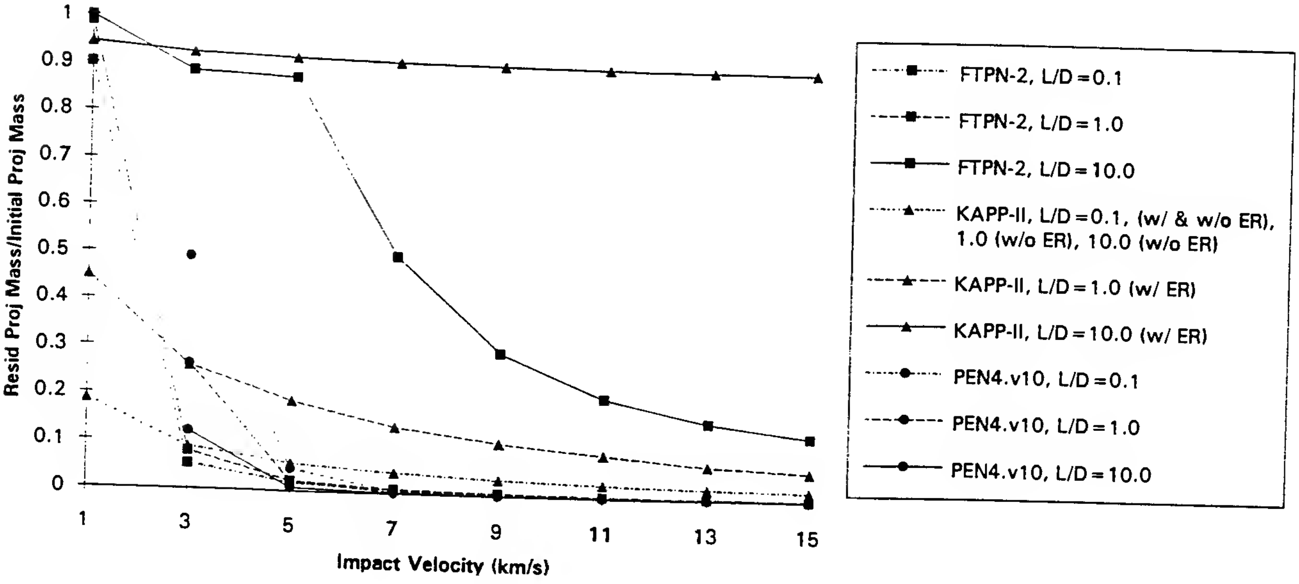
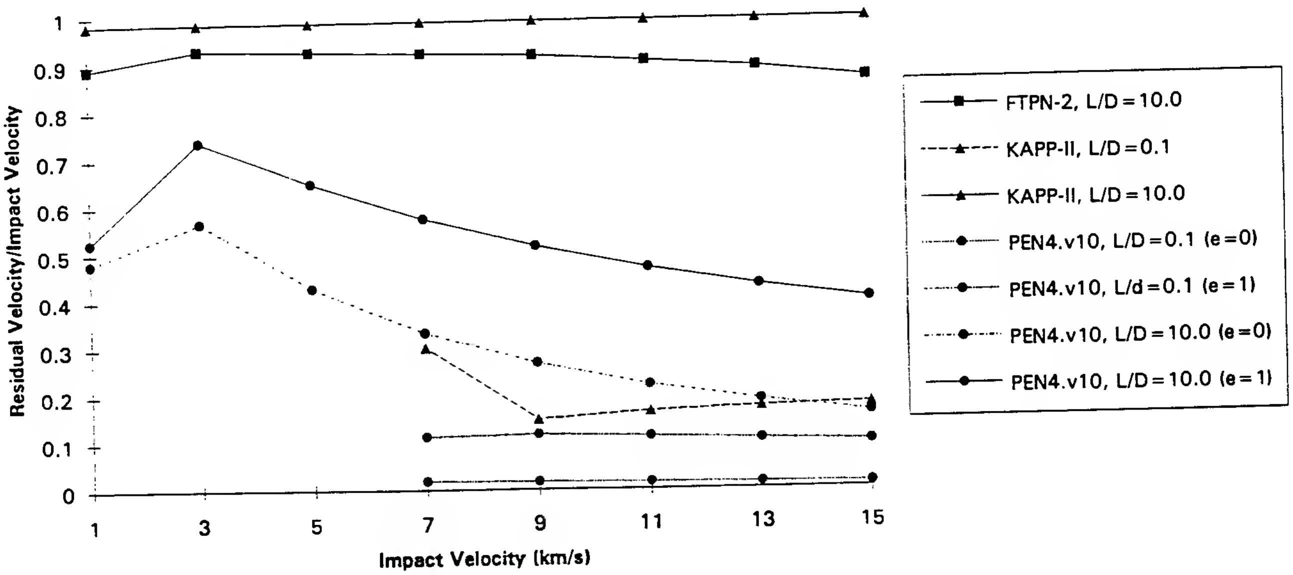


FIGURE 4
RESIDUAL PROJECTILE MASS
ST-->AL, D/T=10.0



## FIGURE 5

RESIDUAL VELOCITY
AL-->AL, D/T=1.0, L/D=0.1 & 10.0



## FIGURE 6

RESIDUAL VELOCITY
AL-->AL, D/T=1.0, L/D=1.0

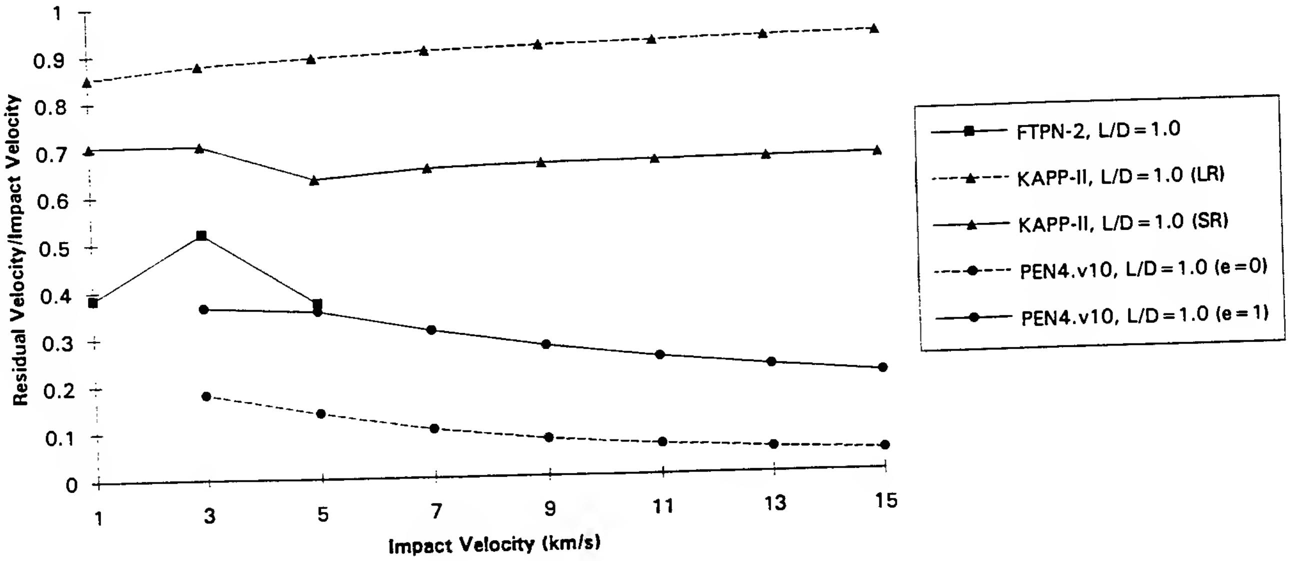


**FIGURE 7**
**RESIDUAL VELOCITY**
**AL-->AL, D/T=10.0, L/D=0.1 & 10.0**

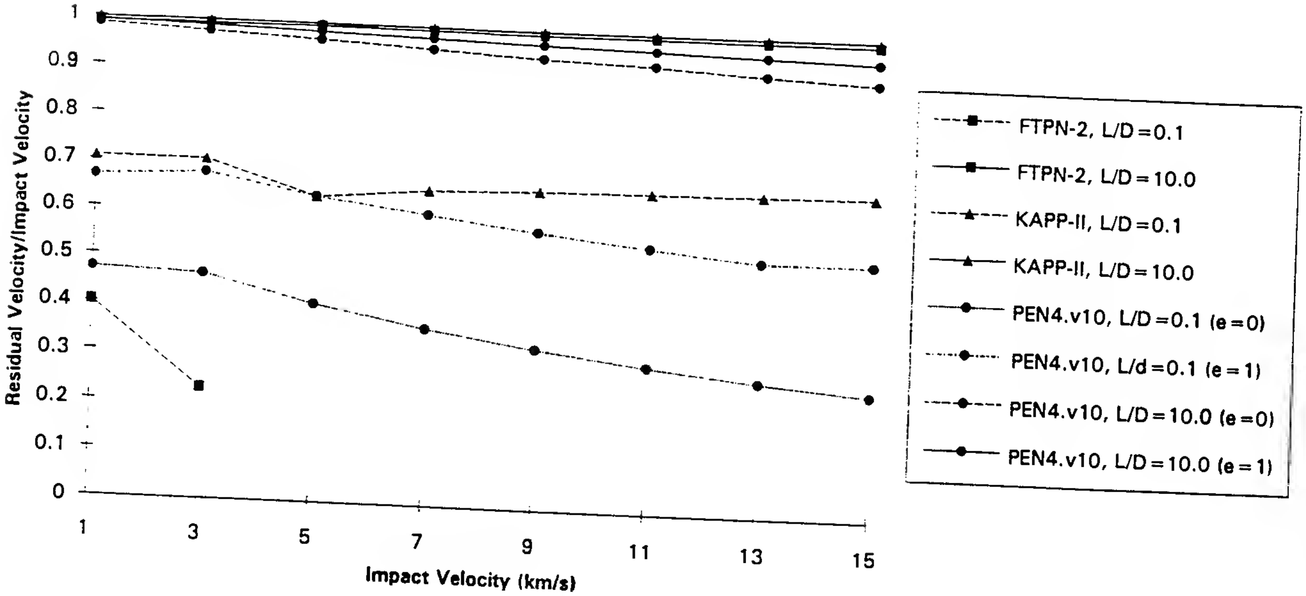


**FIGURE 8**
**RESIDUAL VELOCITY**
**AL-->AL, D/T=10.0, L/D=1.0**

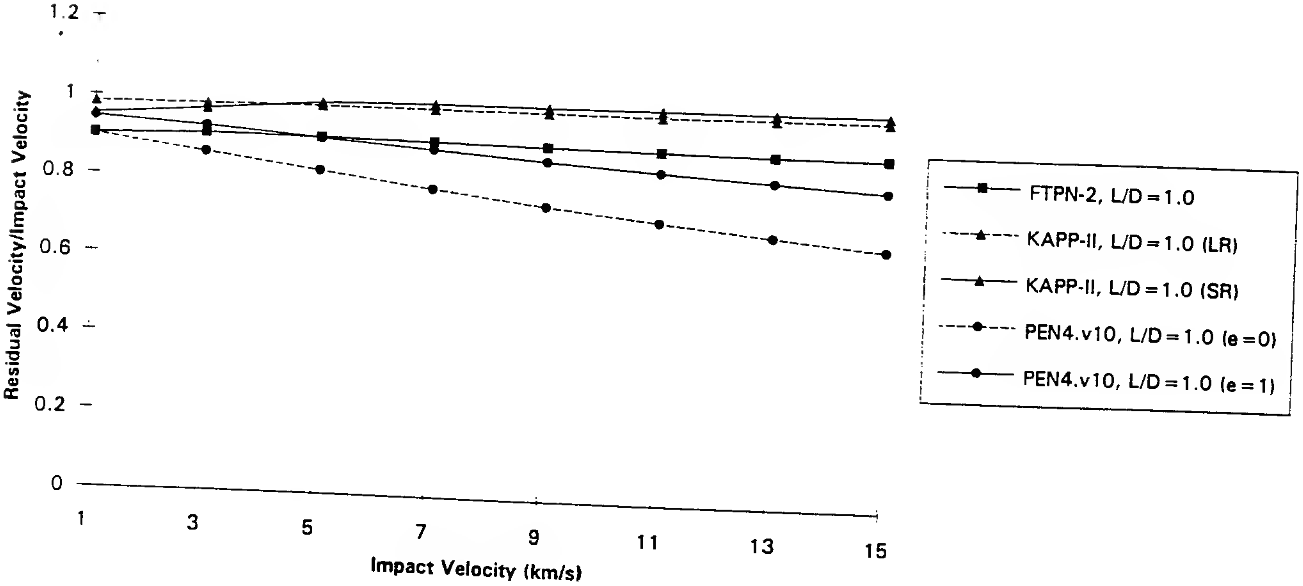


FIGURE 9

RESIDUAL VELOCITY

ST-->AL, D/T=0.1, L/D=0.1 & 10.0

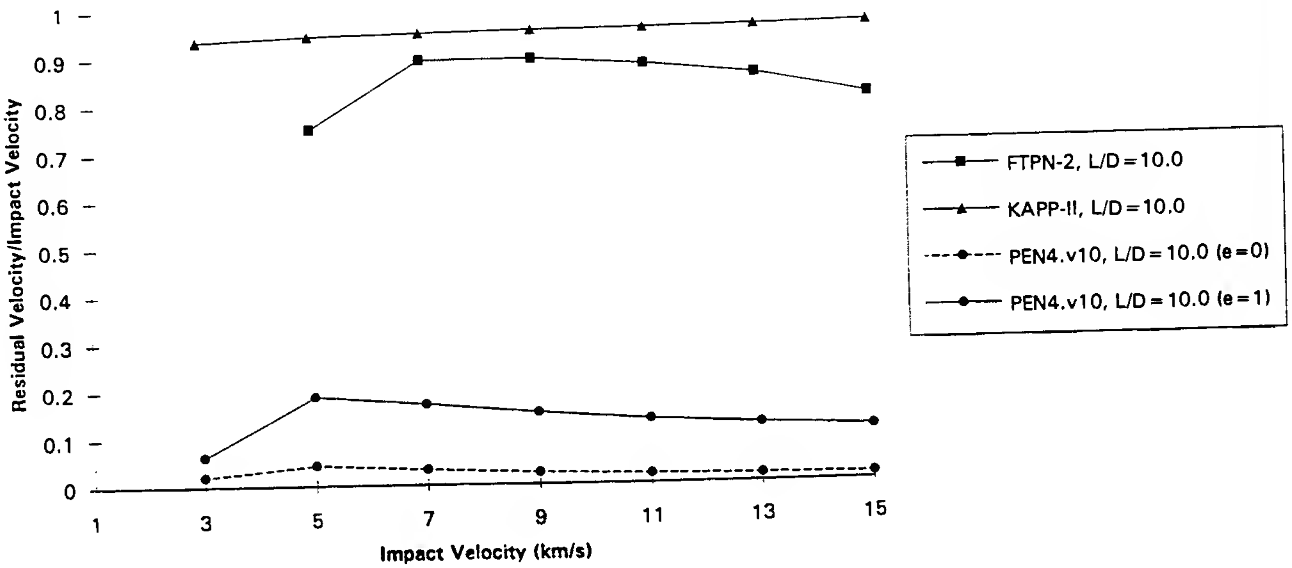


FIGURE 10

RESIDUAL VELOCITY

ST-->AL, D/T=0.1, L/D=1.0

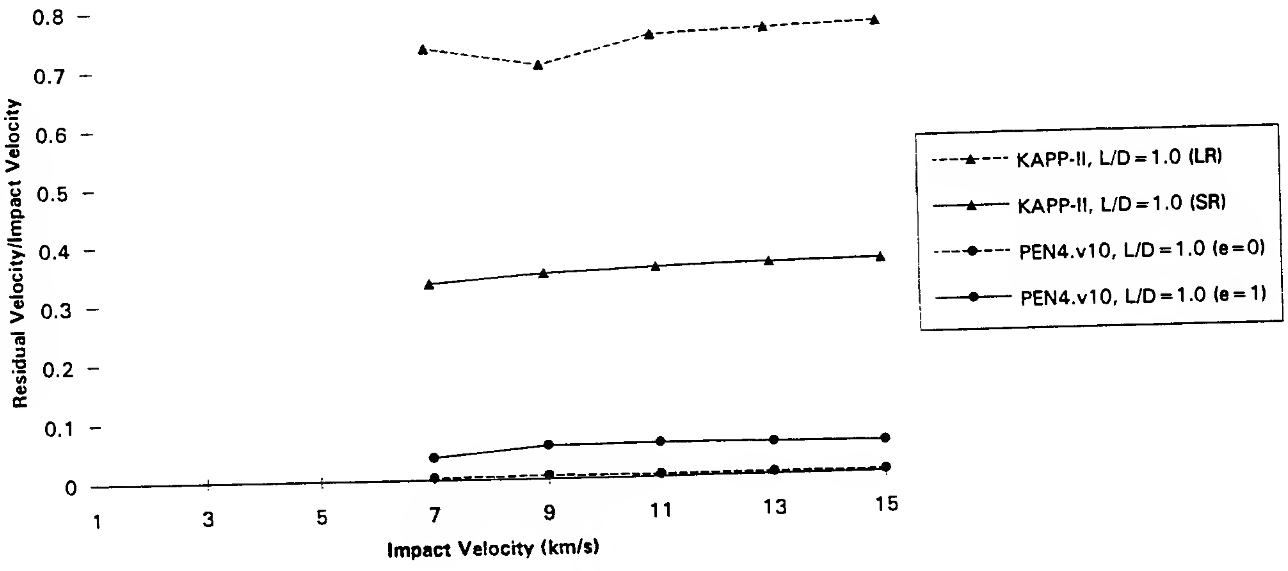


FIGURE 11
RESIDUAL VELOCITY
ST-->AL, D/T=1.0, L/D=0.1 & 10.0

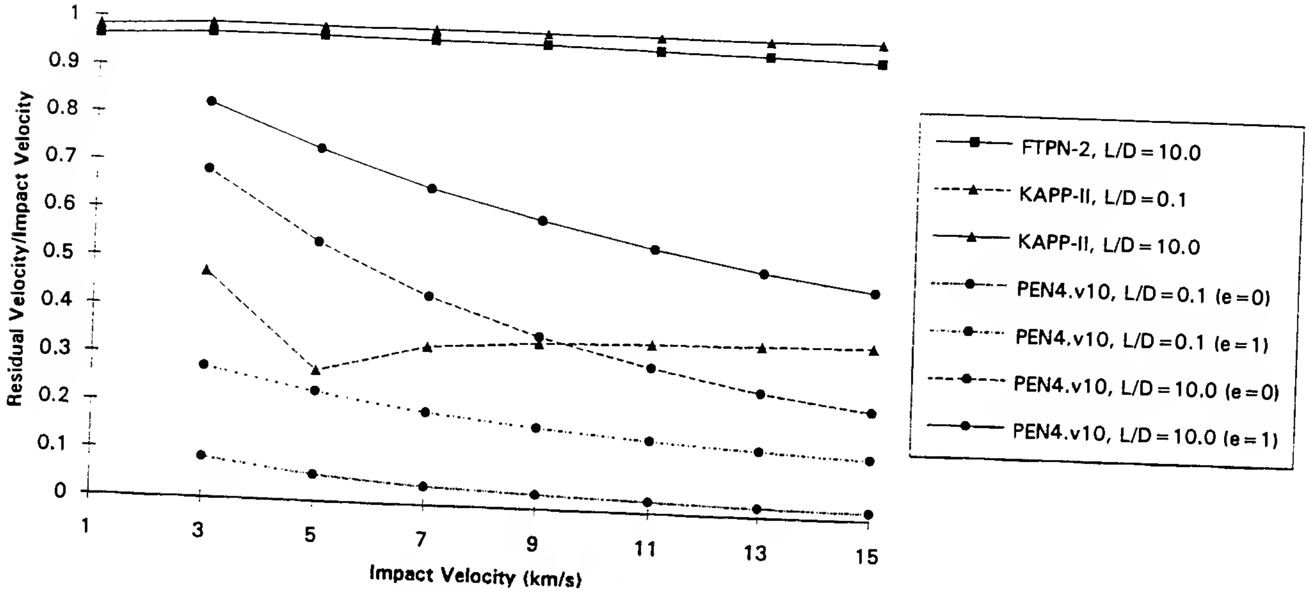
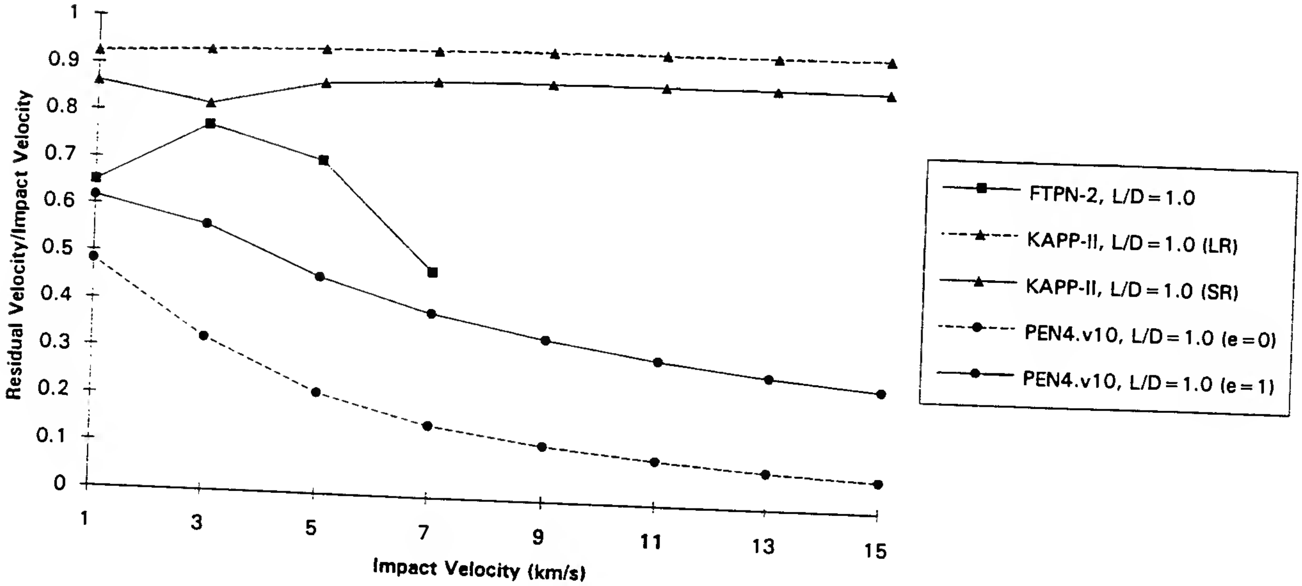
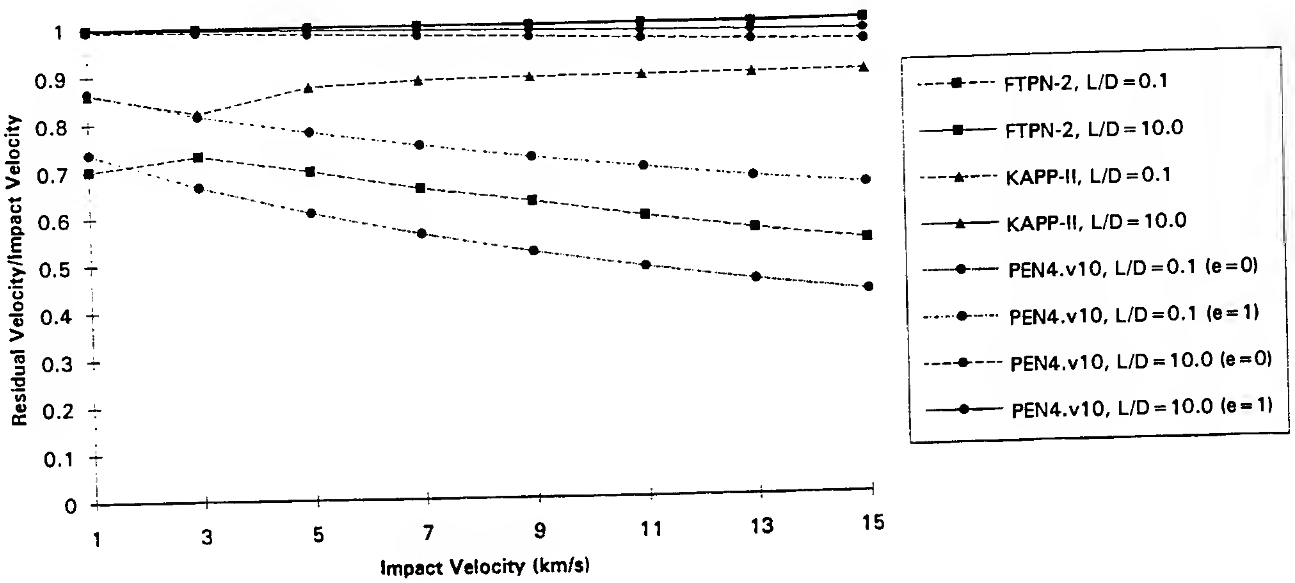


FIGURE 12
RESIDUAL VELOCITY
ST-->AL, D/T=1.0, L/D=1.0



## FIGURE 13
### RESIDUAL VELOCITY
### ST-->AL, D/T=10.0, L/D=0.1 & 10.0



## FIGURE 14
### RESIDUAL VELOCITY
### ST-->AL, D/T=10.0, L/D=1.0

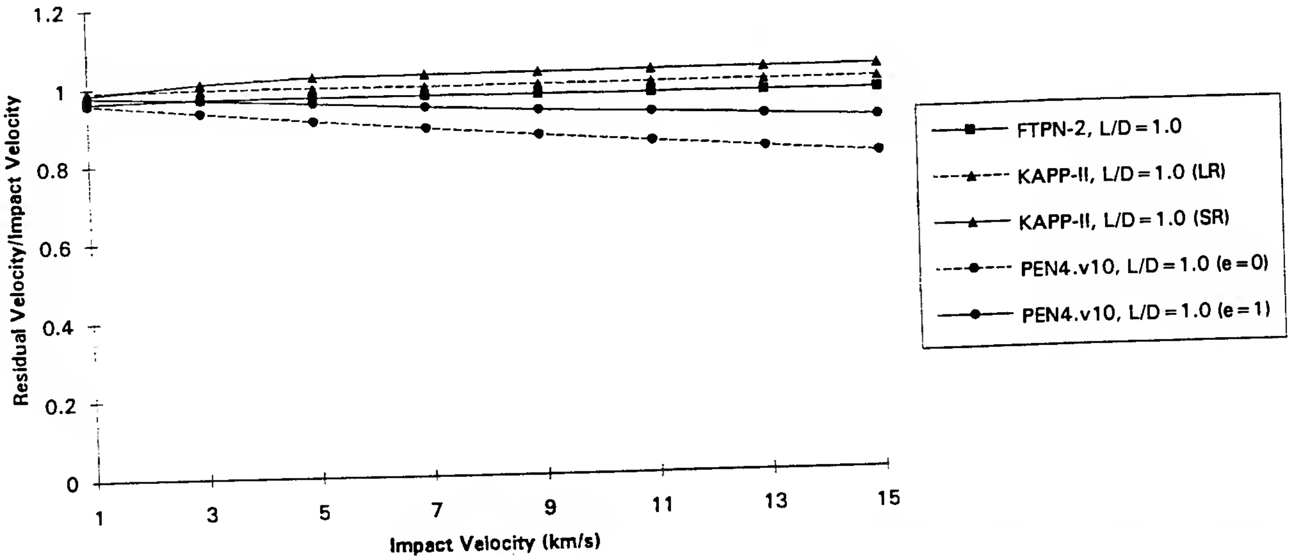


FIGURE 15

NUMBER OF PROJECTILE FRAGMENTS
AL-->AL, D/T=1.0

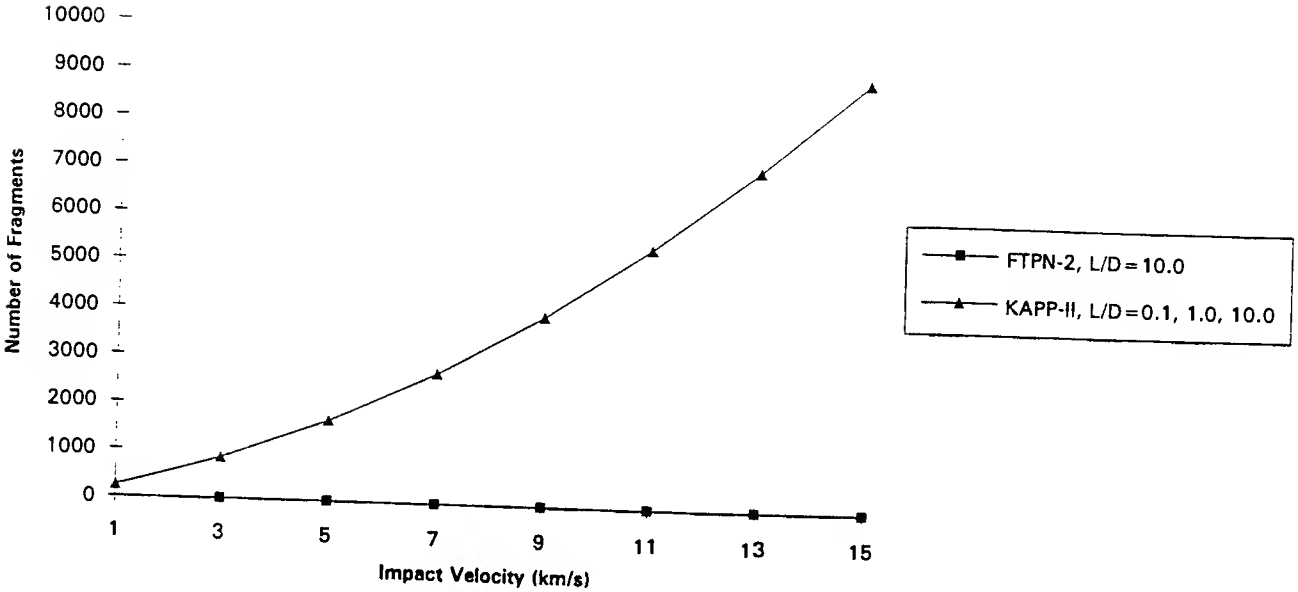


FIGURE 16

NUMBER OF PROJECTILE FRAGMENTS
AL-->AL, D/T=10.0

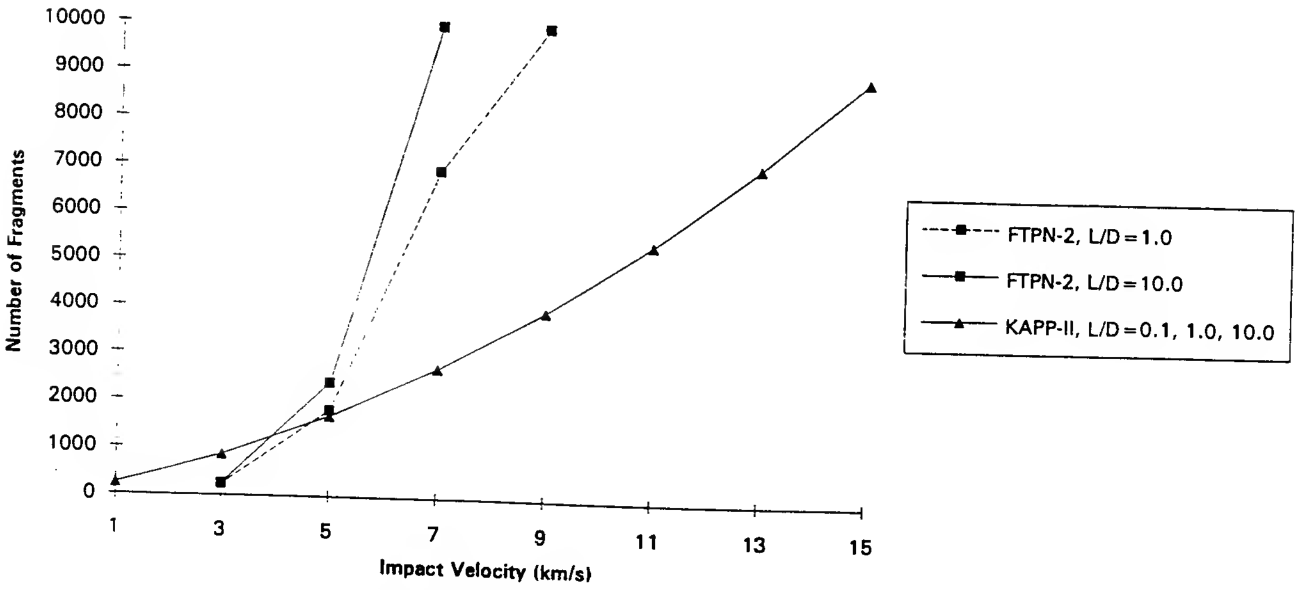


FIGURE 17
NUMBER OF TARGET FRAGMENTS
AL--> AL, D/T = 1.0

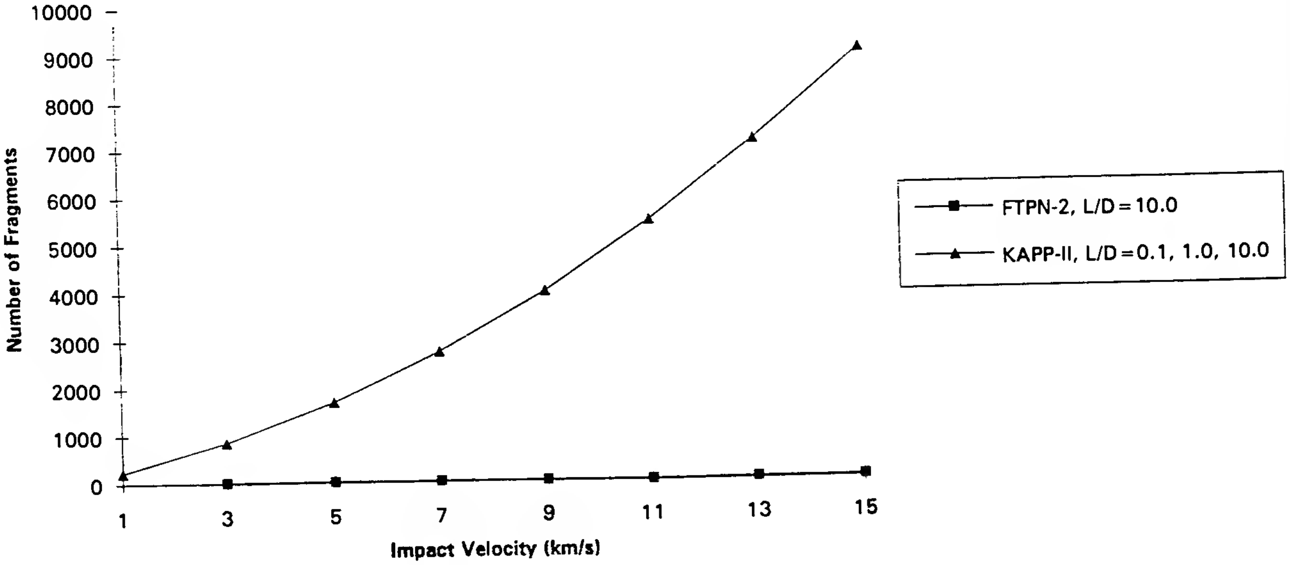


FIGURE 18
NUMBER OF TARGET FRAGMENTS
AL--> AL, D/T = 10.0

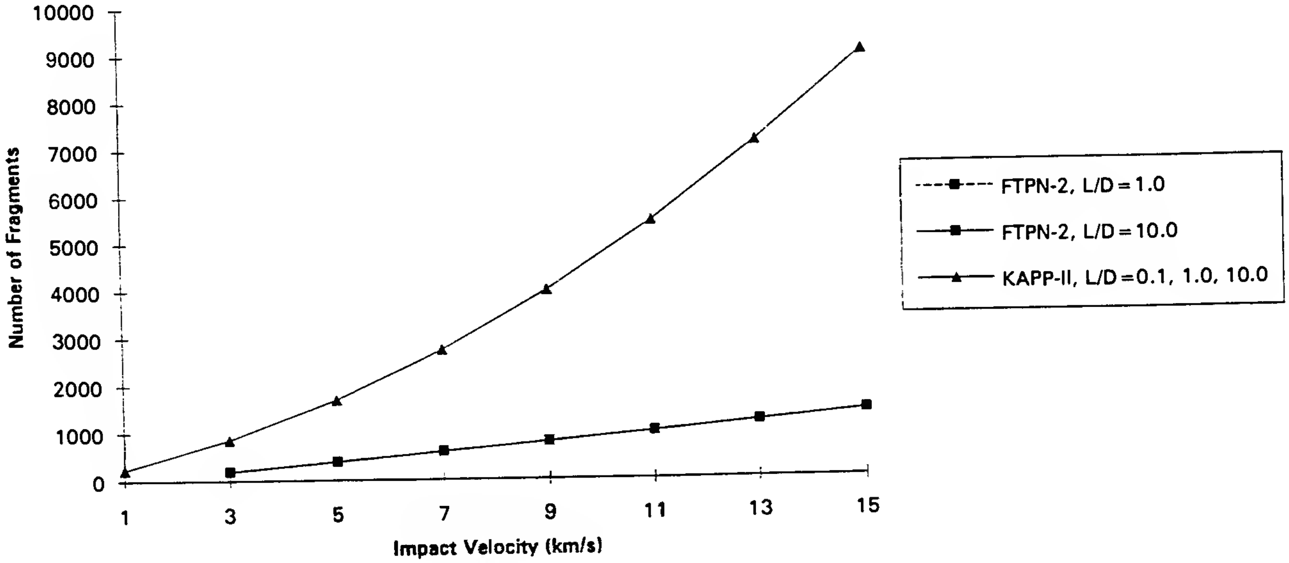


FIGURE 19

NUMBER OF PROJECTILE FRAGMENTS
ST-->AL, D/T=0.1

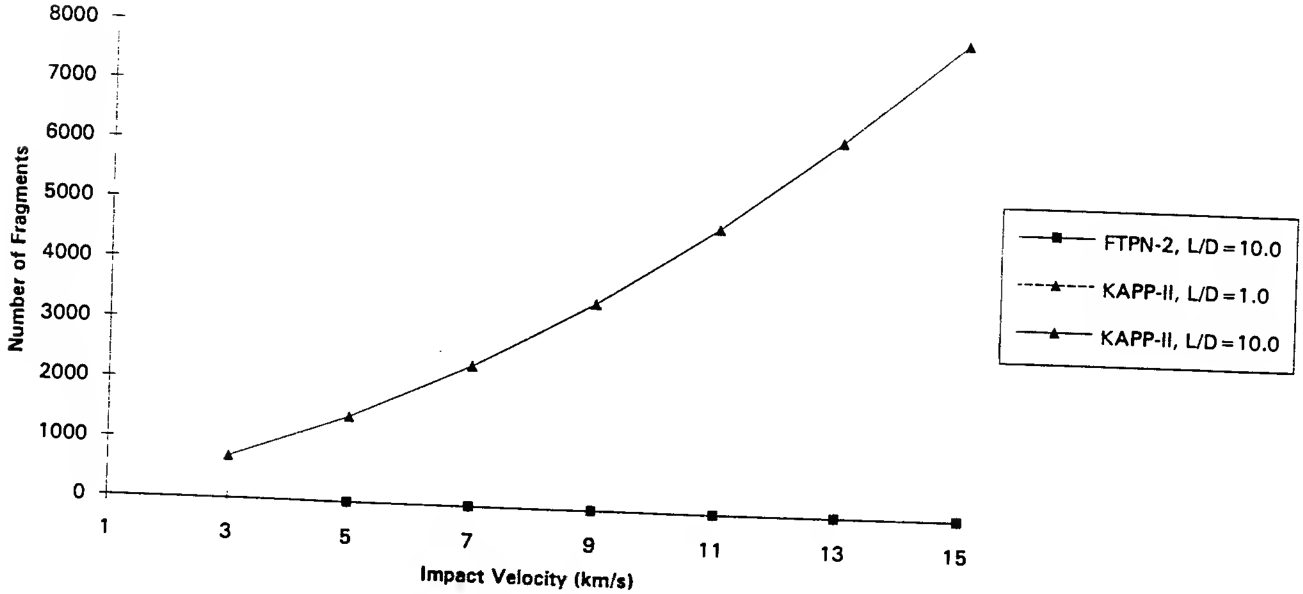
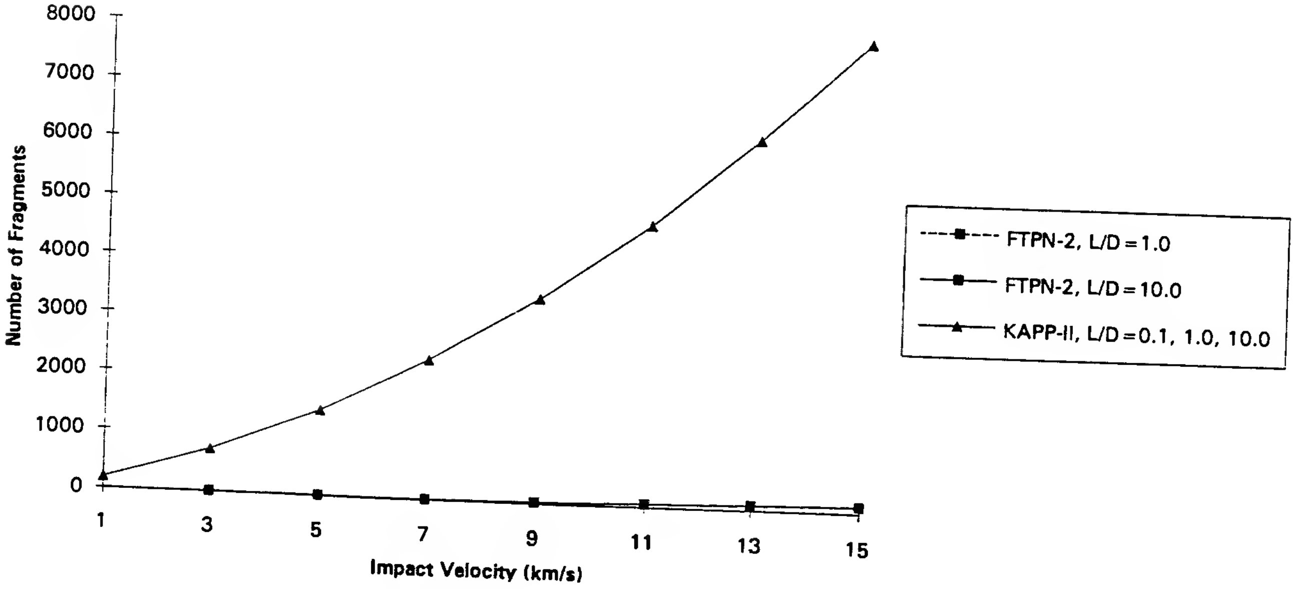
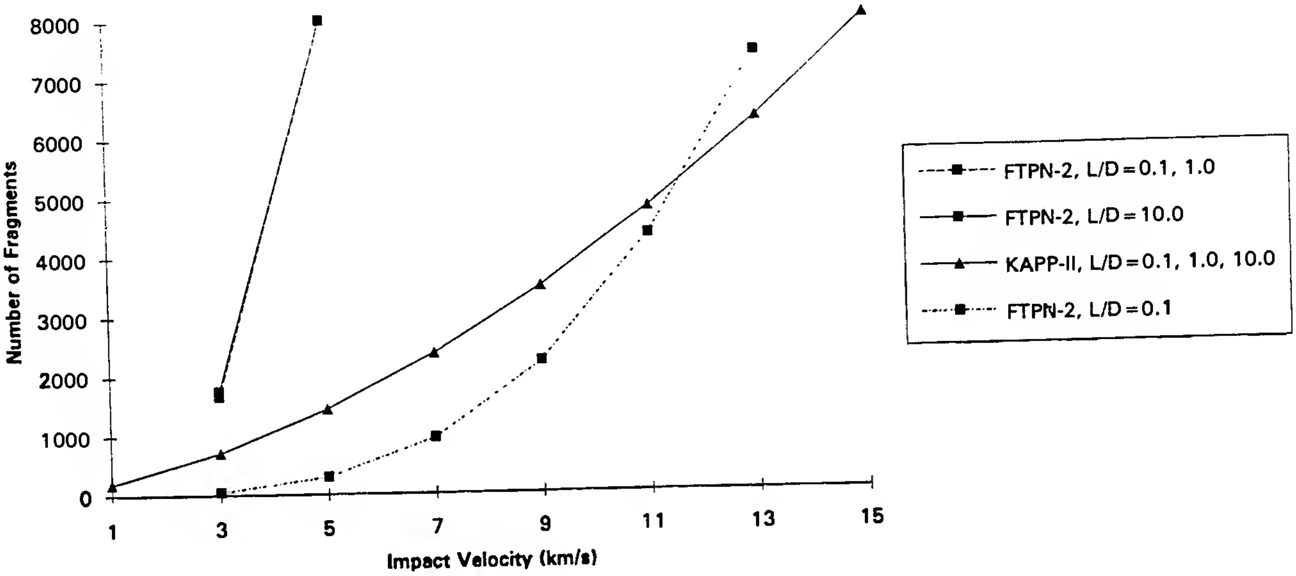


FIGURE 20

NUMBER OF PROJECTILE FRAGMENTS
ST-->AL, D/T=1.0



## FIGURE 21
NUMBER OF PROJECTILE FRAGMENTS
ST-->AL, D/T=10.0



## FIGURE 22
NUMBER OF TARGET FRAGMENTS
ST-->AL, D/T=0.1

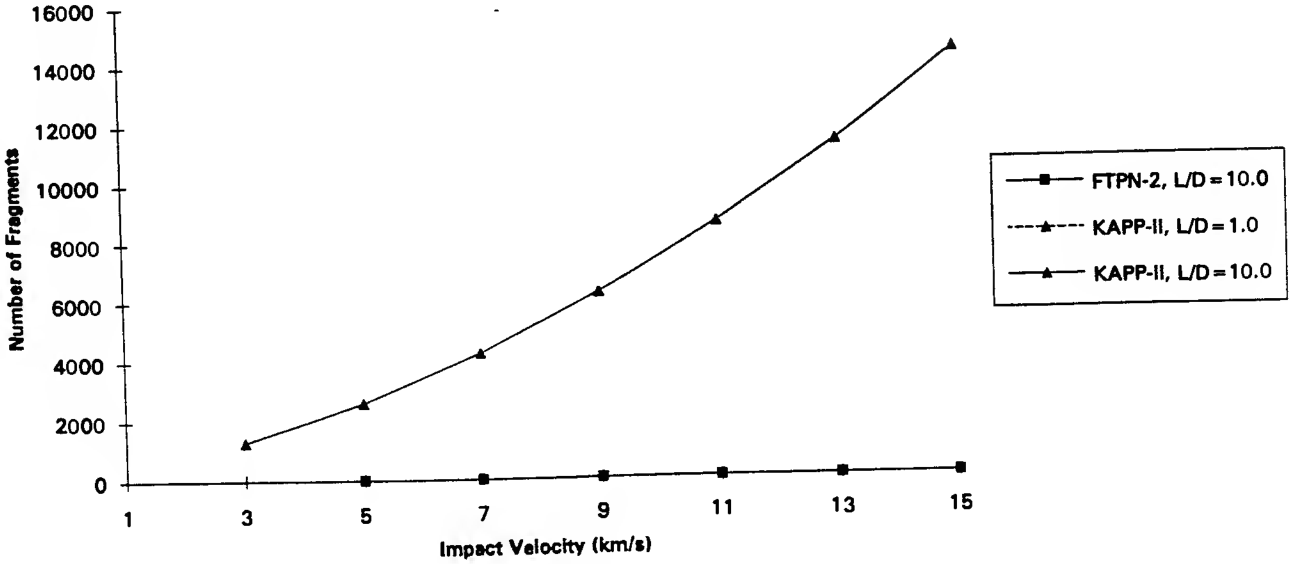


FIGURE 23

NUMBER OF TARGET FRAGMENTS
ST-->AL, D/T=1.0

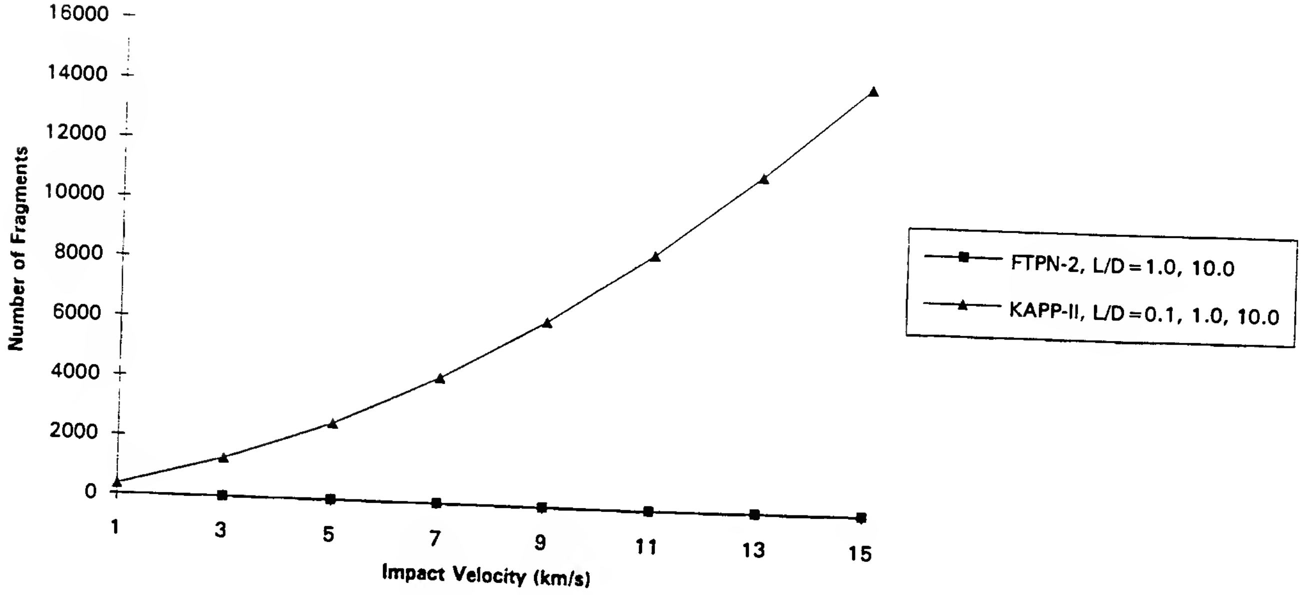


FIGURE 24

NUMBER OF TARGET FRAGMENTS
ST-->AL, D/T=10.0

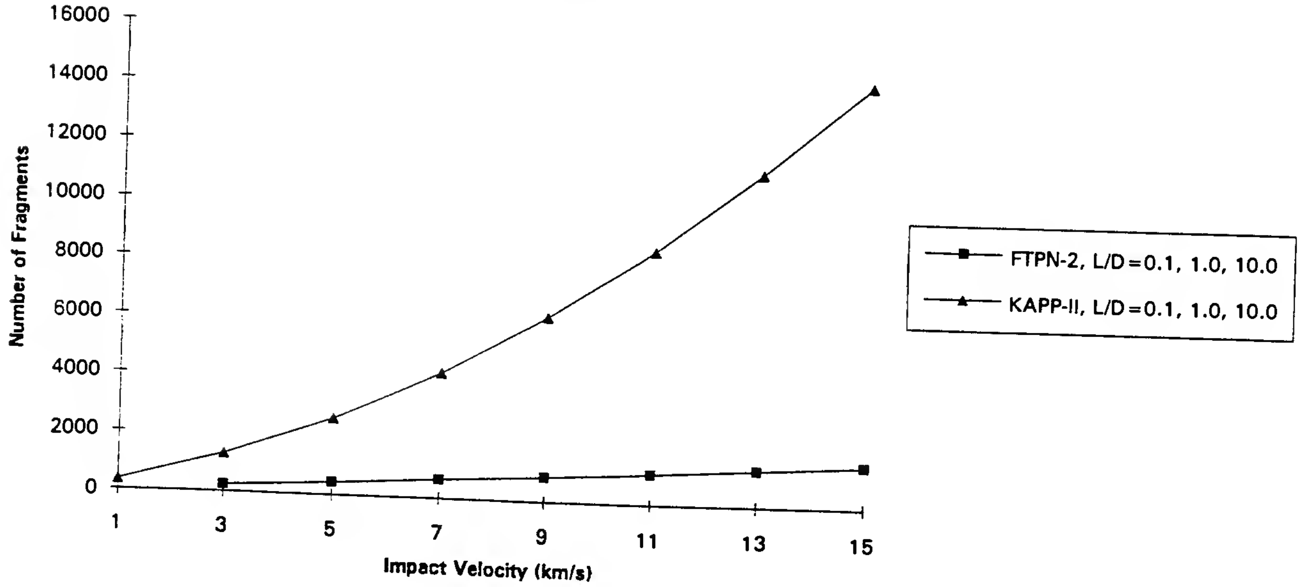


FIGURE 25

DEBRIS CLOUD HALF-ANGLE
AL-->AL, D/T=1.0

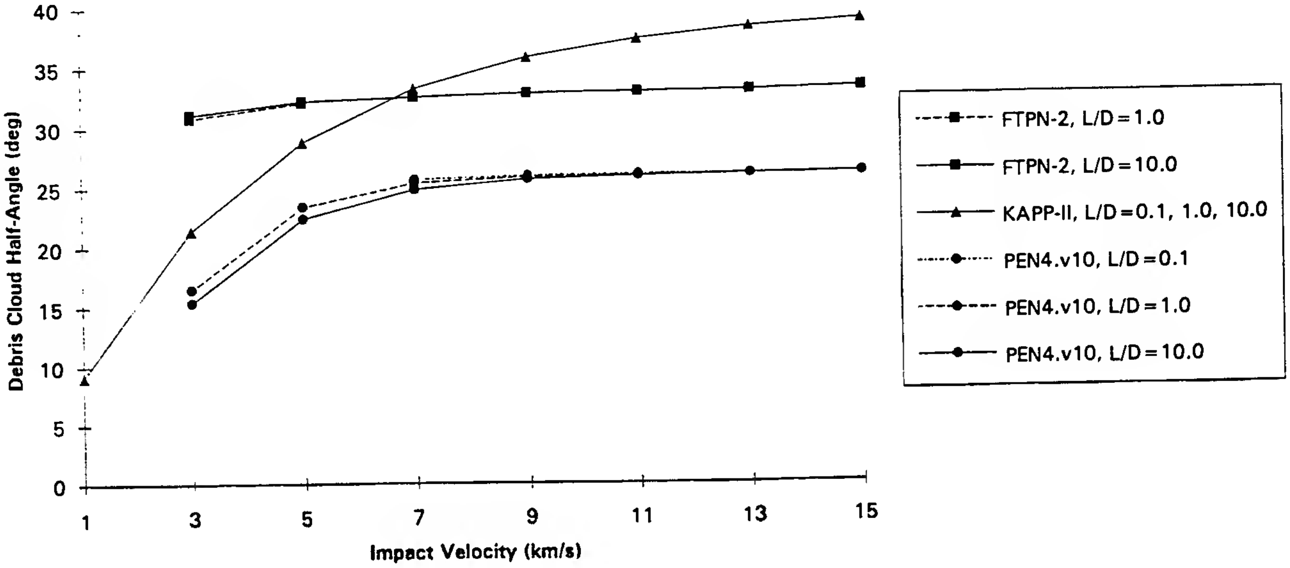


FIGURE 26

DEBRIS CLOUD HALF-ANGLE
AL-->AL, D/T=10.0

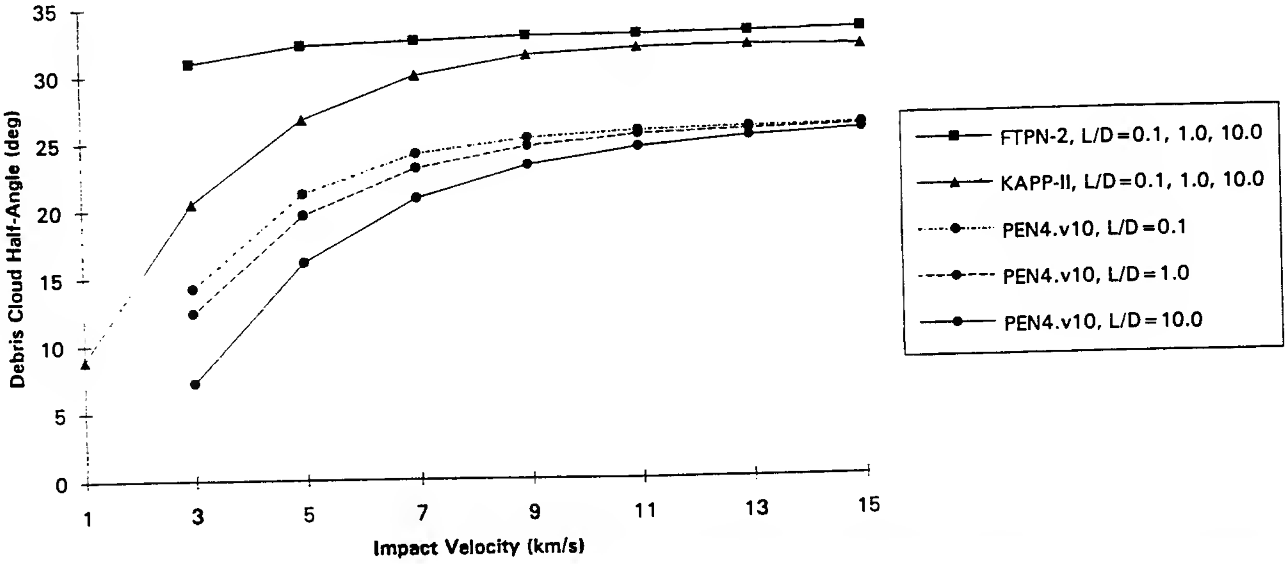


FIGURE 27
DEBRIS CLOUD HALF-ANGLE
ST-->AL, D/T=0.1

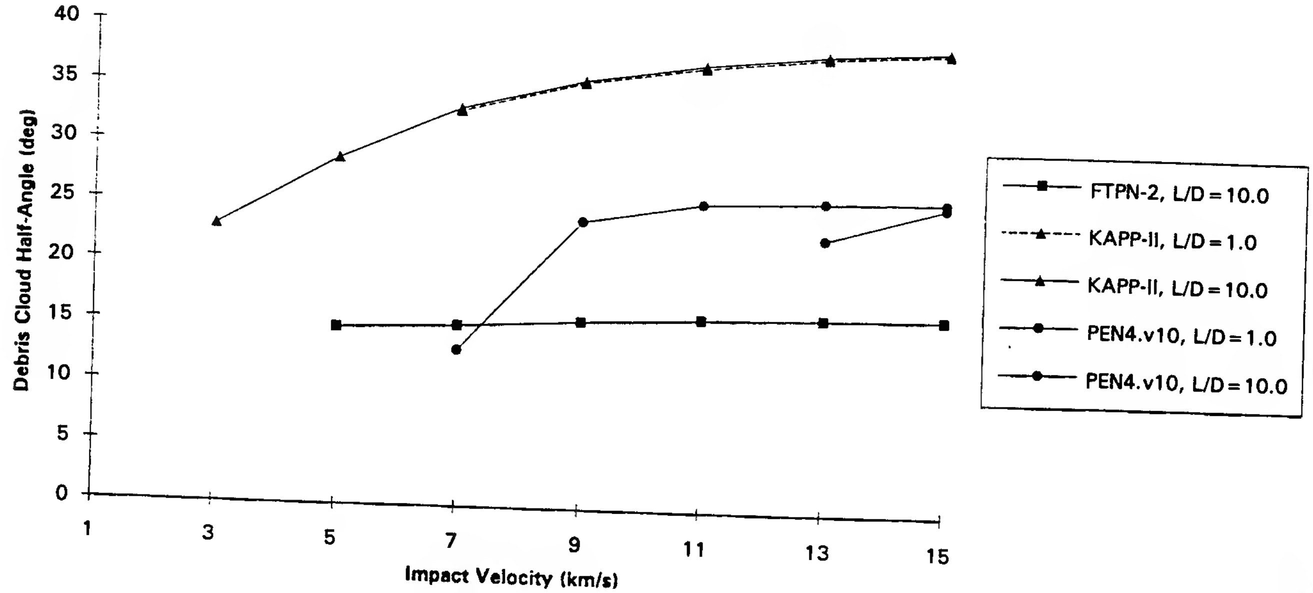


FIGURE 28
DEBRIS CLOUD HALF-ANGLE
ST-->AL, D/T=1.0

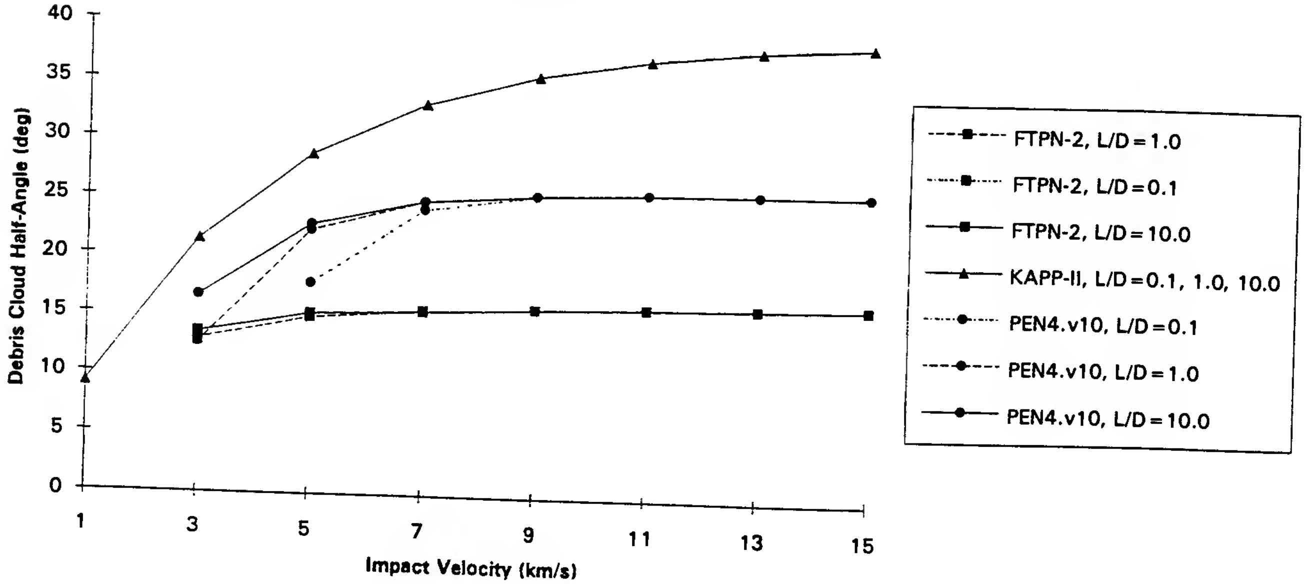
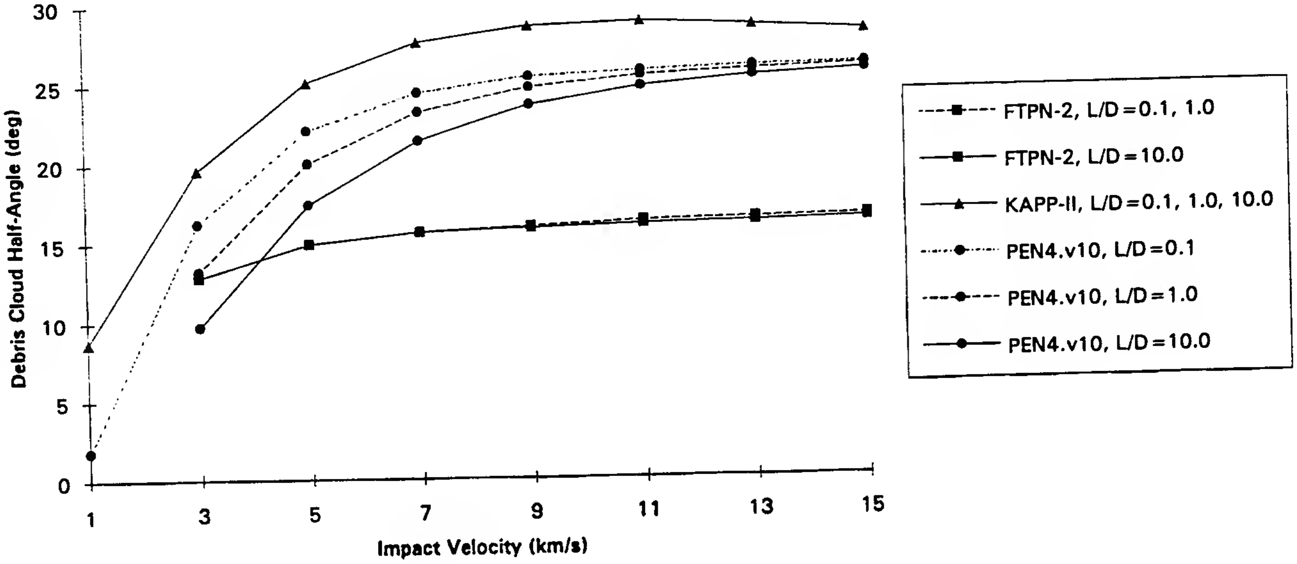


FIGURE 29

DEBRIS CLOUD HALF-ANGLE
ST-->AL, D/T=10.0



# Digital Signal Processing Algorithms for Digital EW Receivers

Faculty Associate : Arnab K. Shaw
Assistant Professor
Electrical Engineering Department
Wright State University
Dayton, OHIO-45435

Final Report for :
Research Initiation Program
Wright Laboratory, Avionics Directorate

Sponsored By :
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D. C.

December 1993

# Digital Signal Processing Algorithms for Digital EW Receivers

Arnab K. Shaw
Assistant Professor
Electrical Engineering Department
Wright State University

## Abstract

In this project, the major focus was to develop and study signal processing algorithms for estimation and detection of parameters useful for localization and identification of targets in Electronic Warfare (EW) environment. Two algorithms for improved estimation of Angles-of-Arrival (AOA) and radio frequencies (RF) were considered.

One of the major contributions of this work is the development of an efficient method for estimating AOA/RF without any eigendecomposition or iterative optimization. Presently, the Minimum-Norm method (MNM) for high-resolution Angles-of-Arrival (AOA) estimation relies on special-purpose hardware or software for obtaining the signal and noise subspace eigenvectors of Autocorrelation (AC) matrices [1, 2, 50, 51]. It is shown in this report that the DFT of the AC matrix (DFT-of-AC) essentially performs an equivalent task of separating the signal and noise subspaces. Furthermore, when the signal-subspace part of the DFT-of-AC vectors are used in the MNM framework, almost identical high-resolution AOA estimates are produced.

Next, a Maximum-Likelihood Estimator (MLE) that ensures unit circle frequencies is presented. A recently proposed MLE approach (KiSS-IQML) converts the frequency estimation problem into a problem of estimating the coefficients of a $z$-polynomial with roots at the desired frequencies [5, 30, 31, 55]. Theoretically, the roots of the estimated polynomial should fall right on the unit circle. But KiSS-IQML, as originally proposed, do not guarantee that. This drawback sometimes causes merged frequency estimates, especially at low SNR [31, 55]. If all the sufficient conditions for the entire $z$-polynomial to have unit circle roots are incorporated, the optimization problem becomes too nonlinear and it loses the desirable weighted-quadratic structure. A novel approach is introduced in this report, where the exact constraints are imposed on each of the 1st-order factors corresponding to individual frequencies. The constraints are applied during optimization *alternately* for each frequency. In absence of any merged frequency estimates, the RMS values approach closer to the theoretical Cramer-Rao (CR) bounds at low SNR levels.

# Digital Signal Processing Algorithms for Digital EW Receivers

**Arnab K. Shaw**

## I : Introduction

Digital processing of microwave signals in Electronic Warfare (EW) environment poses a great challenge to researchers in Signal Processing. EW receivers are used for passive localization and identification of target radars. All microwave receivers used in practice utilize analog signal processing techniques [49, 65-68]. The frequency-band of the EW signals are in the GHz range and the signals have wide bandwidths which necessitate sampling and processing of a massive amount of data at or near real-time. No presently existing EW receiver process microwave radar signals entirely in the digital domain. But it is expected that with the emergence of increasingly faster and inexpensive digital computers and high-speed A/D converters, digital processing of microwave signals would most certainly be the way of the future.

The Electronic Support Measures (ESM) group at the Avionics Directorate, WPAFB, Ohio, has been engaged in researching the digital microwave receiver design problem for the past few years. Their work has resulted in many patents and important publications [37, 49, 65-68]. The proposal author had the opportunity to collaborate with the researchers of the ESM group as a summer research faculty in June-August, 1992 and also during the period covered by the present Research Initiation Project. The major goals of the summer research were to develop and implement Digital Signal processing algorithms in order to study their effectiveness and usefulness to the existing Digital Receiver Program. The following algorithms have been studied :

- *Time-Domain Detection of targets in the presence of noise* : The time-domain detection problem has been considered for single and multiple samples. Detection thresholds and Probability of Detection based on Neyman-Pearson Criterion have been derived and coded.
- *Adaptive Frequency/AOA estimation* : The effectiveness of the Direct-Adaptive-Frequency-Estimation (DAFE) algorithm [5] has been studied for realistic signal/noise conditions. Some modifications of DAFE have also been incorporated for improving its performance.
- *Prony's Algorithm* : This well-known algorithm's effectiveness was studied with signals obtained after passing through a hard-limiter (to avoid A/D saturation) and quantizer.
- *Minimum Norm Method Without Eigendecomposition* : One of the major contributions of this work is the development of an efficient method for estimating AOA/RF without any eigendecomposition or iterative optimization. It is shown in this report that the DFT of the AC matrix (DFT-of-AC) essentially performs the task of separating the signal and noise subspaces. Furthermore, when the signal-subspace

part of the DFT-of-AC vectors are used in MNM, almost identical high-resolution AOA estimates are produced.

- *Maximum Likelehood Method With Exact Constraints* : A recently proposed class of Maximum Likelihood algorithms (MLE), referred to as KiSS/IQML, estimate the frequencies or AOAs from the roots of $z$−polynomials [5, 30, 31, 55]. But the estimated roots are not guaranteed to fall on the unit circle, as desired. Based on the theoretical results on zeros of polynomials [36], a new approach is proposed here that will ensure unit circle roots.

Details of the first three topics are included in the final report submitted at completion of the summer research work in October 1992. In this report, the details on the later two topics will be given. In fact, further work on all these topics has been conducted and the focal point at WPAFB, Dr. James B. Y. Tsui has been briefed periodically on the progress of our research.

The purpose of this follow-up research was to complement and continue the work initiated in summer. Some of the work conducted for this project extend our previous work. But we have also developed new algorithms which may have direct application in the digital receiver design effort. Furthermore, the theoretical work presented here are of more general nature and should also find wider applications. We hope that the intermediate goal of the proposed research, which is to develop advanced signal processing algorithms that can effectively exploit the capabilities offered by newly emerging digital technologies, has been achieved to a large extent. Further research on our ultimate aim of building a digital microwave receiver prototype, is presently being conducted with support from the AFOSR.

### I.1 : The Digital Microwave Receiver Design Problem

The primary task of a microwave receiver is to gather data for sorting of signals and for identifying the radar-type. Based on these information, jamming, weapon delivery or other decisions are considered. In order to perform these tasks, the receiver must analyze the received radar pulses and measure or estimate the following six parameters : Angle-of-Arrival (AOA), Radio Frequency (RF), Time of Arrival (TOA), Pulse Amplitude (PA), Pulse Width (PW) and Polarization (P). These parameters may be useful in more than one stages of receiver operation. For example, AOA, RF, TOA, PW and Polarization data are used for signal sorting, whereas the RF, TOA, PW and P are utilized for source identification purposes. PA threshold may also be used for detecting the presence of any source signal. For jamming or subsequent weapon delivery all these parameters need to be analyzed properly.

Unlike most conventional radars the EW receiver design problem is complicated by the fact that no knowledge about the input signal is available to the receiver. The nature of the problem also requires that measurements and decisions be taken immediately or within a few seconds in an entirely passive mode [49, 65-68]. Furthermore, in order to reduce search time and the consequent response time, the processing bandwidth must also be as wide as possible. It is also desirable to have high sensitivity and large dynamic range such that a broad range of signals, including weak ones, can be detected.

### I.2 : Background and Motivation

In the past two decades, many classes of radar and sonar receivers have been converted from conventional analog technology to purely digital or hybrid systems [24], but EW receivers are yet to make such a transition. The primary technological factors that have been holding back possible fabrication of any digital EW receiver are probably twofold. Firstly, if Analog-to-Digital (A/D) converters are to be used at the operating frequency range, then the Nyquist rate would necessitate sampling at the GHZ range and secondly, the digital hardware or firmware must have the capacity to process such high data rate and produce effective results at or near real-time. But even though the carrier frequencies are in the GHz range, the bandwidths of the useful signals are only in the 10s of MHz. Hence, an obvious compromise in such a situation would be to down-convert the original signal to an intermediate frequency (IF) band before sampling. Down-conversion or superheterodyning is also quite common in analog microwave receivers because it is much easier to design accurate IF amplifiers and filters having fixed and predetermined bandwidths [49, 65-67]. Frequency down-conversion may cause image signals at the IF band and standard cures used in analog superheterodyne receivers, such as the use of I and Q channels and image suppression filters can be utilized to reduce these effects. Furthermore, multirate sampling/processing or sub-band coding may also be useful to avoid high sampling rate.

Digital EW receivers can be expected to offer some major advantages over their analog counterparts. Foremost among these is the almost lossless storage capability of digital memories which can eliminate the dependence on lossy analog delay lines. Digital processors and memory chips are relatively inexpensive, compact in size and low in weight and the trends are towards even further reductions. Digital signal processing algorithms and digital computing technology have matured tremendously and offer a wide range of capabilities. Parallel processing, pipelining, RISC, VLSI design, systolic architecture, vectorization and array processing, fault tolerant computing and etc., are only some of the well-known aspects of digital computing that the last few decades of research have produced. As our research progresses, we intend to study if some of these ideas can be incorporated in the digital receiver in order to improve the efficiency and accuracy of its performance.

A broad range of digital signal processing algorithms are already available for detection as well as for parametric and non-parametric estimation from observed data [13, 17, 24, 48]. These techniques are based on well-established theory on random processes and applied linear algebra. Among the six parameters noted above, the AOA and frequency information are probably the most important ones for sorting, identification and jamming and a rich body of literature is available for high-resolution AOA and frequency estimation [1-13, 15-35, 37-48, 50-65, 69-79].

Much of the results on AOA/frequency estimation appearing in signal processing literature have been developed for sonar and low-frequency radar applications. These mathematical and statistical theories are mostly valid for the EW scenario. But considering the high data-rate in the present application, computationally simpler algorithms must be developed. One of the major contributions of this work is the development of an efficient method for estimating AOA/RF without any eigendecomposition or iterative optimization.

### I.3 : Historical Perspective on the Research on AOA/Frequency Estimation

Estimation of angles-of-arrivals and radio frequencies pose the greatest difficulty because of the nonlinear nature of the optimization problem. An adaptive estimation scheme developed in [10] is attractive because of its computational simplicity as well as its real-time adaptive capabilities. As part of the research performed this summer, it has been demonstrated that the use of higher estimation model order than the actual order improves the bias and variance of the estimates at the cost of somewhat longer convergence time. For this part of the project, an efficient approach forming signal subspace is considered and an accurate method for Maximum Likelihood estimation of frequencies is presented.

The AOA estimation problem is mathematically equivalent to the Frequency Estimation problem which has been a major research topic in many areas of science. Indeed, in the last couple of hundred years, the search for 'hidden periodicities' from observed data has appeared in varied forms in several seemingly differing disciplines of science. To appreciate the sustained appeal of this problem to researchers over the past two centuries, consider that as far back as in 1795, Prony proposed a simple procedure to estimate the parameters of a multiple sinusoids model of an observation record [16, 44]. But even in modern signal processing literature, useful modifications of Prony's work for noisy data are being reported [28, 69]. About hundred years following Prony's work, Schuster had introduced the idea of periodogram in 1898, while determining the periodicities of meteorological phenomenon [46]. In Schuster's time, the calculation of the periodogram was computationally a very expensive procedure. But with the advent of digital computers and after the discovery of the Fast Fourier Transform (FFT) algorithm by Cooley and Tukey [12], the periodogram has become the standard choice for frequency/AOA estimation in a variety of important applications. The multiple sinusoids model has also been used in radio astronomy for analyzing data received at a finite aperture telescope to resolve the locations of closely spaced stars [4]. It also has wide applications in geophysics, radar, sonar and biological signal analysis and ideas emerging from a large variety of fields have provided a certain maturity to this fundamental problem.

#### I.3.a : The Resolution Limitation of the Periodogram

Ever since its discovery in 1965, the FFT has been the primary tool for estimating Angles of Arrival (AOA) or frequencies of far-field sources from noisy observation data. The software or hardware implementation of FFT is remarkably straight-forward. To date, the periodogram continues to be the most frequently used method for frequency/AOA estimation [40, 46]. In fact, it is well known that for localizing a single target, if the noise in the observed data is Gaussianly distributed, the periodogram [46] produces the maximum likelihood estimate. But in case of multiple targets, the periodogram cannot resolve two frequencies which are separated by less than the bin-width of the FFT. In fact, when the sources are spaced at less than the DFT bin-width, the periodogram fails to distinguish two closely spaced frequencies and only provides a single frequency estimate instead of two. The last statement truly portrays the problem one faces while resolving two closely spaced sinusoids when a relatively short data record is available. Clearly, if any amount of data is available for processing, the periodogram of sufficiently zero-padded data will provide reasonably good estimates. But in many problems of practical interest only short data record is available and one has to overcome the periodogram's resolution limitation by resorting to what are commonly known in the signal processing literature as 'High-Resolution' or 'Superresolution' techniques. The major contributions in the higher resolution approaches are highlighted next.

### I.3.b : High-Resolution Methods

A multitude of AOA/Frequency Estimation algorithms, their variations and analysis are available in the literature [1-13, 15-35, 37-48, 50-65, 69-79]. In the following paragraphs only some of the major developments are briefly discussed.

*Minimum Variance Method* : In order to improve upon Periodogram's resolution limit, Capon had proposed this linear estimator which minimizes the interference at frequencies outside the band of interest [9]. Its performance has been shown to be better than the periodogram estimator but worse than the modeling based estimators [34].

*Model-Based Methods* : A major motivation for many modern high-resolution frequency estimation methods has come from the desire to achieve more exact models for the sinusoids-in-noise data. In the Parameter Estimation area in the theory of Statistics, it had been well established that Auto-Regressive (AR) modeling is very appropriate for modeling data with peaky spectra. But in the frequency estimation field also, it had been a common knowledge that data composed of sinusoids in noise tend to have peaky spectra. Consequently, frequency estimation based on AR-modeling has received considerable attention [7, 8, 18, 23, 35, 38, 44, 48, 72, 73].

Depending on how the autocorrelation values are estimated, there are three types of AR parameter estimation methods, namely, Autocorrelation method [35], Covariance method [35], and Modified Covariance method (also known as the Forward-Backward method) [38, 73]. The later two cases are more appropriate for sinusoidal processes because of their implicit relationship with Prony's method which provides perfect frequency estimates when no noise is present. Incidentally, the Maximum Entropy method proposed by Burg [7, 8] and the Linear Prediction based spectral estimator [35], both produce essentially identical frequency estimates as the Covariance method.

When $p$ sinusoids are present and a $p^{th}$ order AR model is used, the frequency estimates are found to be poor at low SNR ($\leq 30dB$). To circumvent this hurdle, larger order ($L > p$) AR model has been proposed [28, 72]. The larger model order tends to accommodate a major part of the interfering noise and thereby reduces the effect of noise in the estimates. The larger-order approach performs poorly below 20dB SNR [28].

*Eigen-Analysis of the Auto-Correlation Matrix of Sinusoid-in-Noise Data* : Since the mid-to-late seventies, a whole new class of algorithms are being developed by effective exploitation of the special properties of the autocorrelation matrix of the sinusoids-in-noise data. For $N = p + 1$, the eigendecomposition of $\mathbf{C}$ was first utilized by Pisarenko [42] who showed that the $z$-polynomial formed with elements of the eigenvector corresponding to the smallest eigenvalue has roots at the signal frequencies. Though the idea is elegant, Pisarenko's method performs quite poorly for noisy signals. Pisarenko's approach was later improved upon by Kumaresan [28] where, for $N > p$ cases, all the noise eigenvectors had been utilized. As an alternate approach, it was shown in [28, 29] that the signal subspace eigenvectors can also be utilized to form a noise subspace vector which should have zeros at the signal frequency locations. This was achieved in [28, 29, 45] by formulating a Minimum-Norm criterion which is the framework that will be used in the proposed work.

Another major improvement on Pisarenko's approach was presented by Schmidt [50, 51] and Bienvenue and Kopp [1, 2]. They proposed to combine the eigenvectors corresponding to the $(L-p)$ smaller eigenvalues of $\mathbf{C}$ and used an orthogonality criterion to obtain the frequency estimates. In the literature, this approach is known as the 'MUSIC' method.

It may be pertinent to emphasize here that the approach proposed in this work for extracting signal or noise subspace 'without eigendecomposition' may be combined with either the MNM or the MUSIC framework. The MNM framework has been preferred in developing the DFT-based MNM (D-MNM) because in case of the Minimum-Norm method, the frequencies are found directly from the polynomial roots. On the other hand, a search procedure is necessary in case of MUSIC for estimating the frequencies. The polynomial version of MUSIC, known as 'root-MUSIC', could also be used but in that case the order of the $z$-polynomial would be twice that of MNM.

*Maximum-Likelihood Method* : This class of algorithms maximize the likelihood function for the observed data, leading to optimization of a non-linear criterion which can only be performed iteratively. Several different approaches are available in the literature [5, 30-32, 46, 47, 54-56, 61, 78] and among these the Constrained MLE approach described in this report appears to offer the most accurate results [61].

*Other Methods and Importance of the Proposed Methods* : As listed in the references, there are a large number other methods that address the high-resolution Frequency/AOA estimation problems. In order to achieve the desired high-resolution capability, all these algorithms utilize some form of eigenanalysis or non-linear optimization, both of which are computationally intensive for real-time applications. Among the two high-resolution methods described in this report, the first method is more efficient than the existing Eigen-based methods whereas the second provides the most accurate frequency estimates.

# II : High-Resolution Angle of Arrival (AOA) Estimation Without Eigendecomposition

## II.1 : Introduction

The Fast Fourier Transform (FFT) is an efficient technique for calculating Discrete Fourier Transform (DFT) at uniformly spaced bins. In many important practical applications, such as radar, sonar and astronomy etc., the resolution capability of FFT is inadequate. Overcoming the resolution limitation of DFT has been a vigorously researched topic in Signal Processing in the past three decades. The modern methods attain the desired 'High-Resolution' or 'Superresolution' at the cost of steep computational burden. The existing well-known methods utilize Eigen-Decomposition (ED), Singular Value Decomposition (SVD) or Maximum Likelihood (ML) computation or nonlinear optimization. These algorithms can only be implemented iteratively which limits their real-time capabilities.

The primary objective for this part of the project is to study whether the computational simplicity of DFT can be effectively combined with the underlying mathematical framework of high-resolution methods. The desired goal is to achieve high-resolution without any iterative optimization. Specifically, some well-known existing approaches, such as the Minimum-Norm method (MNM), extract the signal and noise subspace information from the eigenvectors of the Autocorrelation (AC) matrices. It is shown that the DFT of the AC-matrix (DFT-of-AC) essentially performs an equivalent task of extracting and decoupling the signal and noise subspace information. Hence, it is proposed that the signal eigenvectors be replaced by the largest-norm DFT-of-AC vectors. It is demonstrated that when the DFT-of-AC vectors with larger norms are used in the MNM framework, mostly better or almost equivalent high-resolution AOA estimates are produced. The bias, mean-squared error and the root locations of the proposed DFT-based-MNM (D-MNM) compare well with the Eigendecomposition-based MNM (E-MNM). The simulations further show that the performance of the D-MNM is more robust at low SNR and it has superior dynamic range. The major significance of the proposed work is that, no complicated iterative optimization is needed and the signal-subspace information is extracted only by a *single matrix multiplication.* Hence, hardware implementation of D-MNM for real-time high-resolution AOA/Frequency estimation may be feasible with currently available technology.

## II.2 : Problem Definition

This part of the project addresses the problem of estimating of the Angles of Arrival (AOA) of densely spaced narrowband targets. Suppose that $p$ plane waves originating from far-field point sources at distinct directions impinge on a linear array of $N$ equally spaced sensors. The signal sampled simultaneously at $m^{th}$ instant of time at $N$ equally spaced sensors form a 'snapshot' vector defined as,

$$\mathbf{x}_m \triangleq [x_m(0)\ x_m(1)\ \ldots\ x_m(N-1)]^T. \qquad (II.1)$$

In the presence of noise, the observation samples can be written as,

$$x_m(n) = \tilde{x}_m(n) + z_m(n) \qquad (II.2)$$

where, $z_m(n)$ represents the additive observation noise and/or the modeling error and $\tilde{x}_m(n)$ denotes the signal part of the observation, which is given by

$$\tilde{x}_m(n) = \sum_{i=1}^{p} A_m(i) e^{j\frac{2\pi d}{\lambda}(n - \frac{N+1}{2})\sin\theta_i + j\phi_m(i)}, \qquad n = 0, 1, \ldots, N-1 \qquad (II.3)$$

where,

$p$ : Number of narrowband sources present

$d$ : Spacing between sensor elements

$\lambda$ : Wavelength of radiation of the received signals

$\theta_i$ : Angles-of-Arrival (AOA) of the $i^{th}$ source

$A_m(i)$ : Amplitude of the $i^{th}$ source at the $m^{th}$ snapshot

$\phi_m(i)$ : Phase angle of the $i^{th}$ source at the $m^{th}$ snapshot,
Uniformly distributed between $-\pi$ and $\pi$.

The noise $z_m(n)$ is assumed to be zero-mean and uncorrelated with the source signals and it has a variance of $\sigma_z^2$. The signal model can be written in a more succinct form as,

$$\tilde{x}_m(n) = \sum_{i=1}^{p} A_{im} e^{j\omega_i n} \qquad (II.4)$$

where, $\omega_i$ and $A_{im}$ are defined as

$$\omega_i \triangleq \frac{2\pi d}{\lambda} \sin\theta_i \qquad \text{and} \qquad (II.5)$$

$$A_{im} \triangleq A_m(i) e^{-j\frac{2\pi d}{\lambda}\left(\frac{N+1}{2}\right)\sin\theta_i + j\phi_m(i)}. \qquad (II.6)$$

Further details about the above model may be found in [13]. With the above formulation the model for the observation matrix can be written as,

$$\tilde{\mathbf{X}} \triangleq \mathbf{TA} \qquad (II.7)$$

where,

$$\mathbf{T} \triangleq \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & \cdots & e^{j\omega_p} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\omega_1(N-1)} & e^{j\omega_2(N-1)} & \cdots & e^{j\omega_p(N-1)} \end{bmatrix}, \qquad (II.8)$$

$$\triangleq [\mathbf{t}_1 \ \mathbf{t}_2 \ \cdots \ \mathbf{t}_p], \qquad (II.9)$$

$$\mathbf{A} \triangleq [\mathbf{a}_1 \ \mathbf{a}_2 \ \ldots \ \mathbf{a}_M] \qquad \text{and} \qquad (II.10)$$

$$\mathbf{a}_m \triangleq \begin{bmatrix} A_{1m} \\ A_{2m} \\ \vdots \\ A_{pm} \end{bmatrix} \qquad \text{for } m = 1, 2, \ldots, M. \qquad (II.11)$$

For half wavelength spacing between two successive sensors of the line array, $\omega_i = \pi \sin\theta_i$. With $M$ snapshot vectors defined in ($II$.2), the $N \times M$ observation matrix $\mathbf{X}$ is formed as,

$$\mathbf{X} \triangleq [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_M]. \tag{II.12}$$

Using the observation matrix, the spatial covariance matrix can be estimated as,

$$\mathbf{C} \triangleq \frac{1}{M}(\mathbf{X}\mathbf{X}^H) \tag{II.13a}$$

$$\triangleq \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_m \mathbf{x}_m^H. \tag{II.13b}$$

The description of the observation and the model is now complete. Given the noisy observation matrix $\mathbf{X}$, the problem under consideration in this proposal is to estimate $\omega_i$'s and $A_{im}$'s. Note that the complex amplitudes can be estimated linearly once the $\omega_i$'s are known but the estimation of poses the greatest difficulty because it is a highly nonlinear optimization problem.

The primary objective for this part is to study whether the computational simplicity of DFT can be effectively combined with the underlying mathematical framework of some of the existing high-resolution methods. The final goal is to achieve high-resolution without any iterative optimization such that real-time implementation may be feasible with existing hardware. The proposed method makes use of the special properties of correlation matrices which are outlined next.

### II.3 : Some Properties of the Autocorrelation Matrix

Since the data described by ($II$.3) is uncorrelated, zero mean WSS process, the $N \times N$ ($N \geq p$) covariance matrix $\mathbf{C}$ will have the following matrix decomposition when there is no observation noise,

$$\mathbf{C} = \mathbf{T}\mathbf{\Sigma}\mathbf{T}^H \tag{II.14}$$

where, $\mathbf{\Sigma} \triangleq diag\ (\sigma_1^2 \ \sigma_2^2 \ \dots \ \sigma_p^2)$ and $\sigma_i^2$ denotes the power of the the $i$-th signal. Note that this ideal $\mathbf{C}$ has rank $p$. In this case, the eigen-decomposition of $\mathbf{C}$ can be written as,

$$\mathbf{C}\mathbf{V} = [\lambda_1 \mathbf{v}_1 \ \cdots \ \lambda_p \mathbf{v}_p \ 0 \ \cdots \ 0] = \mathbf{\Lambda}\mathbf{V} \tag{II.15a}$$

$$\triangleq \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p & 0 & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} | & | & \cdots & | & | & \cdots & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_p & \mathbf{v}_{p+1} & \cdots & \mathbf{v}_N \\ | & | & \cdots & | & | & \cdots & | \end{bmatrix}. \tag{II.15b}$$

For observations with noise as defined in ($II$.3),

$$\mathbf{C} = \mathbf{T}\mathbf{\Sigma}\mathbf{T}^H + \sigma_z^2\mathbf{I}. \tag{II.16}$$

Note that this theoretical $\mathbf{C}$ has rank $N$ though the signal part, $\mathbf{T\Sigma T}^H$ has rank $p$. In this case, the eigen-decomposition of $\mathbf{C}$ can be written as,

$$\mathbf{CV} = [\,(\lambda_1 + \sigma_z^2)\mathbf{v}_1 \quad \cdots \quad (\lambda_p + \sigma_z^2)\mathbf{v}_p \quad \sigma_z^2\mathbf{v}_{p+1} \quad \cdots \quad \sigma_z^2\mathbf{v}_N\,] \qquad (II.17)$$

where, the $\lambda_i$'s and $\sigma_z^2$ represent the signal and noise eigenvalues. But in practice, the eigendecomposition has to be performed on the sample covariance matrix $\mathbf{C}$ as defined in $(II.13)$ and then the noise eigenvalues will not be equal but will be absorbed with the signal eigenvalues also. In that case,

$$\mathbf{CV} = [\,\hat{\lambda}_1\mathbf{v}_1 \quad \cdots \quad \hat{\lambda}_p\mathbf{v}_p \quad \hat{\lambda}_{p+1}\mathbf{v}_{p+1} \quad \cdots \quad \hat{\lambda}_N\mathbf{v}_N\,] \qquad (II.18)$$

where, the estimated eigenvalues are ordered as, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \; \cdots \hat{\lambda}_N$. The eigenvectors corresponding to the $p$ largest eigenvalues are called the 'signal eigenvectors' which constitute the 'signal-subspace'. All the other $(N-p)$ eigenvectors are known as the 'noise eigenvectors'. Note also that the $p$ 'signal eigenvectors' of $\mathbf{C}$ span the subspace defined by the columns of $\mathbf{T}$ and that they are orthogonal to the 'noise subspace' eigenvectors.

## II.4 : The Proposed DFT-Based Minimum-Norm Method (D-MNM)

As a significant departure from the eigen-based approaches discussed in the previous section, this work advocates that the signal-subspace information be extracted from the DFT-of-AC matrix which can be accomplished with a single matrix multiplication. This will eliminate the need for iterative calculation of eigenvectors which is computationally intensive. The central idea behind the DFT-of-AC matrix is analyzed first.

### II.4.a : Signal and Noise Subspace Extraction from the DFT-of-AC Matrix

Let the DFT matrix be denoted as,

$$\mathbf{D} \triangleq [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_N\,], \qquad (II.19)$$

where, the elements of the $k$-th DFT-vector $\mathbf{e}_k$ is defined as, $\mathbf{e}_k(l) \;=\; e^{j\frac{2\pi}{N}kl}$, for $k, l = 0,1,\,2,\,\ldots,\,N-1$. If the frequencies $\omega_i$s are all on the DFT bins and if there is no observation noise, then in general,

$$\begin{aligned}
\mathbf{f}_k &\triangleq \mathbf{Ce}_k && (II.20a)\\
&= \frac{1}{M}\sum_{m=1}^{M}(\mathbf{x}_m^H\mathbf{e}_k)\mathbf{x}_m, && \text{using } (I.13b) \qquad (II.20b)\\
&= \frac{1}{M}\sum_{m=1}^{M}(\mathbf{a}_m^H\mathbf{T}^H\mathbf{e}_k)\mathbf{x}_m, && \text{using } (II.8) \qquad (II.20c)\\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbf{a}_m^H\begin{bmatrix}\mathbf{t}_1^H\mathbf{e}_k\\ \vdots \\ \mathbf{t}_p^H\mathbf{e}_k\end{bmatrix}\mathbf{x}_m. && (II.20d)
\end{aligned}$$

If the $k$-th DFT vector $\mathbf{e}_k$ corresponds to one of the $\omega_i$ frequencies,

$$\begin{aligned}\mathbf{f}_k &= \frac{1}{M}\sum_{m=1}^{M} A_{km}^{*}\mathbf{T}\mathbf{a}_m = \mathbf{T}\frac{1}{M}\sum_{m=1}^{M} A_{km}^{*}\mathbf{a}_m \\ &= \mathbf{T}\begin{bmatrix} \frac{1}{M}\sum_{m=1}^{M} A_{km}^{*}A_{1m} \\ \vdots \\ \frac{1}{M}\sum_{m=1}^{M} |A_{km}|^2 \\ \vdots \\ \frac{1}{M}\sum_{m=1}^{M} A_{km}^{*}A_{pm} \end{bmatrix} = \mathbf{T}\begin{bmatrix} \hat{\sigma}_{k1} \\ \vdots \\ \hat{\sigma}_k^2 \\ \vdots \\ \hat{\sigma}_{kp} \end{bmatrix} \end{aligned} \qquad (II.21)$$

where, $\hat{\sigma}_{kl}$s denote the covariance of the complex amplitudes. Assuming the number of samples $M$ to be large and since $A_{km}$s are independent random variables, $\hat{\sigma}_{A_{km},A_{lm}} \triangleq \hat{\sigma}_{kl} = \delta_{kl}\hat{\sigma}_k^2$. Hence,

$$\mathbf{f}_k \rightarrow \hat{\sigma}_k^2\mathbf{t}_k = \hat{\sigma}_k^2\mathbf{e}_k. \qquad (II.22)$$

Note that the norm of $\mathbf{f}_k$ is directly proportional to the signal power, $\hat{\sigma}_k^2$, *i.e.*, this norm will be large if the signal power is significant. On the other hand, if a DFT-vector $\mathbf{e}_k$ does not correspond to any of the $\omega_i$ frequencies then due to orthogonality, $\mathbf{t}_i^H\mathbf{e}_k = 0, \forall i$. For such cases,

$$\mathbf{f}_k = \mathbf{0}. \qquad (II.23)$$

For this ideal case then, the DFT-of-AC has the following decomposition,

$$\begin{aligned}\mathbf{F} &\triangleq \mathbf{C}\mathbf{D} && (II.24a) \\ &\triangleq [\mathbf{f}_1 \quad \mathbf{f}_2 \quad \cdots \quad \mathbf{f}_N] && (II.24b) \\ &\rightarrow [\Lambda_1\mathbf{u}_1 \quad \cdots \quad \Lambda_p\mathbf{u}_p \quad 0 \quad \cdots \quad 0] && (II.24c)\end{aligned}$$

where, the $\Lambda_i$s and $\mathbf{u}_i$s are the lengths and unit vectors of each $\mathbf{f}_i$, respectively. Note that the unit vectors in the matrix in ($II$.24$c$) have been rearranged so that the zero/nonzero components are clustered together. Interestingly, this decomposition appears to be very similar to the usual Eigendecomposition of noiseless and ideal $\mathbf{C}$, as given by ($II$.15). For this ideal signal scenario again, if the DFT-of-AC is formed using the theoretical and noisy Covariance matrix of ($II$.16), then the decomposition has the form,

$$\begin{aligned}\mathbf{F} &= \mathbf{C}\mathbf{D} && (II.25a) \\ &= \mathbf{T}\Sigma\mathbf{T}^H\mathbf{D} + \sigma_z^2\mathbf{D} && (II.25b) \\ &\rightarrow [(\Lambda_1+\sigma_z^2)\mathbf{u}_1 \quad \cdots \quad (\Lambda_p+\sigma_z^2)\mathbf{u}_p \quad \sigma_z^2\mathbf{u}_{p+1} \quad \cdots \quad \sigma_z^2\mathbf{u}_N], && (II.25c)\end{aligned}$$

where the $\mathbf{u}_i$'s have been arranged in decreasing order of lengths. Note again that this decomposition is analogous to the one in ($II$.17). In this case also, the $p$ largest-norm vectors of the DFT-of-AC matrix contain the signal subspace information.

In practice, the $\omega_i$s will not be on the DFT bins and the observations may also be noisy and hence, the decomposition in $(II.24)$ or $(II.25)$ will not hold. But the DFT-components ($\mathbf{f}_k$s) closer to the signal frequencies will tend to have larger norms (this is further analyzed in Section II.6). Hence, for the general scenario, when the observation data is noisy and the angular frequencies $\omega_i$s are arbitrarily spaced, the signal/noise subspace decomposition can be formed as :

$$\mathbf{F} \rightarrow \ [\Lambda_1 \mathbf{u}_1 \ \cdots \ \Lambda_p \mathbf{u}_p \ \mid \ \Lambda_{p+1} \mathbf{u}_{p+1} \ \cdots \ \Lambda_N \mathbf{u}_N] \qquad (II.26a)$$

$$\triangleq \mathbf{\Lambda} [\mathbf{U}_S \ \mid \ \mathbf{U}_N] \qquad (II.26b)$$

where, $\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_N$ are the norms of the $\mathbf{f}_i$ vectors and the matrices $\mathbf{\Lambda}$, $\mathbf{U}_S$ and $\mathbf{U}_N$ are formed as,

$$\mathbf{\Lambda} \triangleq \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_p \end{bmatrix}, \quad \mathbf{U}_S \triangleq \begin{bmatrix} | & | & \cdots & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p \\ | & | & \cdots & | \end{bmatrix} \quad \text{and,} \quad \mathbf{U}_N \triangleq \begin{bmatrix} | & \cdots & | \\ \mathbf{u}_{p+1} & \cdots & \mathbf{u}_N \\ | & \cdots & | \end{bmatrix}. \qquad (II.26c)$$

It may be observed again that the decomposition in $(II.26)$ is analogous to the eigen-based counterpart in $(II.18)$. It may be noted here that in case of the ideal signal cases of $(II.24)$ and $(II.25)$, an unit vector $\mathbf{u}_i$ corresponds to one of the DFT-vector $\mathbf{e}_k$, but in the general case of $(II.26)$, they are linear combinations of the DFT-components close to the signal frequencies (the general case is further analyzed in Section II.6).

### II.4.b : Incorporation of DFT-Based Signal Subspace in Minimum-Norm Framework

The principal idea behind the Minimum-Norm method is to form an appropriate 'noise-subspace' vector $\mathbf{d}$ which is orthogonal to the 'signal-subspace' defined by $\mathbf{U}_S$. Let,

$$D(z) \triangleq \sum_{k=0}^{N-1} d_k z^{-k} \qquad (II.27)$$

be an $(N-1)$-th order $z$-polynomial with $p$ zeros at, $z_k = e^{j\omega_k}$, for, $k = 1, \ldots, p$, corresponding to the AOAs. The coefficient vector is denoted as,

$$\mathbf{d} \triangleq [d_0 \ \ d_1 \ \ \cdots \ \ d_{N-1}]^T, \qquad (II.28)$$

where, $d_0 = 1$. According to the MNM philosophy [26], if $\mathbf{U}_S$ does constitute of the signal-subspace, then $\mathbf{d}$ must be orthogonal to $\mathbf{U}_S$, *i.e.*,

$$\mathbf{U}_S^H \mathbf{d} = \mathbf{0}. \qquad (II.29)$$

$\mathbf{d}$ needs to be found by solving this underdetermined set of equations which has infinite number of solutions. According to [28, 29], the solution that also minimizes the norm $||\mathbf{d}||^2$, possesses the desirable property that all its roots fall inside the unit circle. This 'minimum-norm' solution of $\mathbf{d}$ for solving $(II.29)$ can be expressed as :

$$\mathbf{d} = \begin{bmatrix} 1 \\ -------- \\ -\ \mathbf{G}^H(\mathbf{G}\mathbf{G}^H)^{-1}\mathbf{g} \end{bmatrix}, \qquad (II.30a)$$

where, $\mathbf{U}_S^H$ is partitioned as,

$$\mathbf{U}_S^H \triangleq [\mathbf{g} \;\mid\; \mathbf{G}]. \qquad (II.30b)$$

Once $\mathbf{d}$ is estimated, the $p$ roots of $D(z)$ closest to the unit circle are used to find the AOAs. It may be recalled that in E-MNM the signal-subspace eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$, as defined in $(II.18)$ are used to form $\mathbf{U}_S$ [28, 29, 45]. But in case of the proposed approach, no eigendecomposition is necessary. Post-multiplication of $\mathbf{C}$ by the DFT-matrix $\mathbf{D}$ is all that is required to extract the signal subspace in $(II.26)$.

### II.4.c : Summary of the Proposed D-MNM Algorithm

The key steps and some alternative possibilities are summarized in this Section.

#### II.4.c.1 : Algorithm Steps

1. Form the Covariance Matrix estimate using forward-backward method [28, 29] :

$$\hat{\mathbf{C}} \triangleq \frac{1}{2M}\sum_{m=1}^{M} \mathbf{x}_m\mathbf{x}_m^H + \mathbf{x}_m^b\mathbf{x}_m^{b\,H}. \qquad (II.31)$$

The 'backward' vector is defined as $\mathbf{x}_m^b \triangleq \mathbf{J}\mathbf{x}_m^*$, where, $\mathbf{J}$ denotes the permutation matrix with 1's at the cross-diagonal entries and $^*$ denotes the complex-conjugate operation.

2. Post-multiply $\mathbf{C}$ by the DFT matrix $\mathbf{D}$ to form the DFT-OF-AC matrix, $\mathbf{F} \triangleq \mathbf{CD}$.

3. Form $\mathbf{U}$ as in $(II.26c)$ using the $p$ unit vectors corresponding to the largest norms. Partition $\mathbf{U}_S$ as in $(II.30b)$.

4. Estimate the $\mathbf{d}$ vector using $(II.30a)$ and form the $D(z)$ polynomial using the elements of $\mathbf{d}$.

5. Find the roots of $D(z)$. Pick the $p$ roots closest to the unit circle to find the desired frequencies/AOAs.

#### II.4.c.2 : Alternate Possibilities

*Steps 2 and 3* : Post-multiplication of the AC-matrix by a DFT-matrix has been used here because the decompositions as described in Section II.4.a appear analogous to eigendecomposition. But it is easy show that identical results can be obtained if the AC-matrix is pre-multiplied by a DFT matrix, *i.e.*, the DFT-of-AC matrix can also be formed alternately as, $\mathbf{F}_1 \triangleq \mathbf{DC}$. In that case, the largest norm row vectors of the DFT-of-AC matrix $\mathbf{F}_1$ must be used to form $\mathbf{U}_S^H$ defined in $(II.31)$.

*Step 4* : This step requires inversion of a matrix of dimension $(N-1) \times (N-1)$. This can be avoided by orthogonalizing the $p$ largest norm vectors in $\mathbf{U}_S$. Let, $\mathbf{U}_S^o$ be the new 'signal-subspace' matrix with the orthonormal set of vectors which can be written in partitioned form as,

$$\mathbf{U}_S^{o\,H} \triangleq [\mathbf{g}_o \;\mid\; \mathbf{G}_o]. \qquad (II.32)$$

With these partitioned matrices, $\mathbf{d}$ can again be found in Step-4 as [28, 29],

$$\mathbf{d} = \begin{bmatrix} 1 \\ \text{--------} \\ -\,\mathbf{G}_o^H\mathbf{g}_o/(1 - \mathbf{g}_o^H\mathbf{g}_o) \end{bmatrix}. \qquad (II.33)$$

It may be mentioned here that in [28, 29], $p$ orthonormal signal eigenvectors were used to form $\mathbf{U}_S$, whereas here $\mathbf{U}_S^o$ is formed by orthogonalizing the $p$ largest norm vectors of the DFT-of-AC matrix.

*Step 5* : This step requires rooting of the $(N-1)$-th order polynomial $D(z)$. Instead, the frequencies may also be found from the peaks of the following minimum-norm pseudo-spectrum [28, 29, 64] :

$$P_{MNM}(e^{j\omega}) \triangleq \frac{1}{|D(e^{j\omega})|^2} \qquad (II.34)$$

## II.5 : Simulation Results

In this Section the performance of D-MNM is compared with some of the existing well-known algorithms using commonly used data sets. For the purpose of simulations, AOA and Frequency Estimation problems are treated separately.

### II.5.a : AOA Estimation

**Simulation 1** : *Two Densely-Spaced Targets of Equal Powers* [62, 63]

Planewaves from $p = 2$ sources with $\theta_1 = 18°$ and $\theta_2 = 22°$ incident on N=8 sensors were modeled as in [29, 30, 32]. The number of snapshots, M=10. Fig. 1 shows the norms of the $\mathbf{f}_i$ vectors for 20 trials at 20dB SNR. The two largest $\Lambda_i$s always appear to be more significant than all the smaller ones. Figures 2a and 2b show the roots of $D(z)$ for 50 independent realizations using D-MNM and E-MNM, respectively. The figures show that the roots in both cases are at almost same locations. Table-1 compares E-MNM and D-MNM in terms of the bias and RMS values with 200 independent trials at different SNR values. The results clearly indicate that the performance of D-MNM is quite close to that of E-MNM, though no Eigendecomposition was required in this case. In fact, D-MNM was found to be somewhat more robust (in terms of successful trials) at low SNR ranges.

**Simulation 2** : *Comparison of Dynamic Range with Two Targets of Unequal Powers* [62, 63]

For this example, two well-separated sources are located at halfway between corresponding DFT-bins with $\theta_1 = 7.1808°$ and $\theta_2 = 61.045°$. The SNR for $\theta_1$ is maintained at a fixed value of 30dB whereas the SNR for the signal at $\theta_2$ is gradually reduced from 30dB to 0dB in steps of 2dB. The mean values with 50 independent trials at each SNR for D-MNM and E-MNM are displayed in Fig. 3. Clearly, D-MNM demonstrates superior dynamic range than E-MNM.

### II.5.b : Frequency Estimation

In this Section, the proposed algorithm is compared with the well-known Tufts-Kumaresan (TK) method [28, 69] and MUSIC method [1, 2, 50, 51] via simulations.

**Simulation 3** : *Comparison of High-Resolution Performance and Threshold Enhancement*

The simulation data is generated using the formula [28, 69] :

$$y(n) \ = \ a_1 e^{j2\pi(0.5)n + j\frac{\pi}{4}} \ + \ a_2 e^{j2\pi(0.52)n} \ + \ w(n), \qquad \text{for, } n = 0,\ 1,\ \ldots,\ M-1 \qquad (II.35)$$

where, $w(n)$ is complex white Gaussian noise with variance $\sigma_w^2$. The number of data samples used is, M=25. This data set has been widely used in the literature for studying the performance of various methods. For this data set, it has been shown in [28, 69] that the TK method performs best when high-order ($L \times L$) covariance matrix with $L = 18$ is used with forward-backward covariance matrix [19, 64]. Five hundred independent noise realizations were used to compare the performance of the proposed method with that of TK method and MUSIC. The mean values for three cases at different SNR values are displayed in Fig. 4. The RMSE results are shown in Fig. 5 along with CR Bound for the frequency at $f_1 = 0.52Hz$. The bias and RMSE at different SNR values are also tabulated in Table 2. Clearly, the proposed method extends the performance threshold closer to the CR bound. Hence the performance of the proposed method approaches that of the Maximum-Likelihood method more closely.

### II.6 : Analysis, Discussion and Directions on Further Research

The results presented so far are quite intriguing and can be expected to have far-reaching consequences on simplifying the present practice of frequency/AOA estimation. The proposed approach of forming signal-subspace using DFT without any eigendecomposition also opens up whole new avenues for further research and, at the same time, poses some unanswered questions. Furthermore, it may be possible to extend and incorporate similar ideas in other closely related problems or to develop more simplified algorithms. The theoretical performance of the method needs to be thoroughly analyzed. The major advantage of the proposed approach is that all the signal-subspaces are obtained with a single matrix multiplication. This step may be performed using FFT which is very efficient for hardware and software implementation. Preliminary analysis of the proposed work and some directions for further research are briefly outlined in this section.

1. **Reduced Computational Complexity and Usefulness in High Sampling-Rate Problems** : The major significance of D-MNM is that its high-resolution capability does not rely on any iterative method or eigendecomposition which is also computed iteratively. The lower computational complexity of D-MNM should be attractive in any general frequency/AOA estimation scenario. But the usefulness of the proposed method should be specially significant in those applications where traditional high-resolution methods are yet to make much inroads due mainly to extremely high sampling rate requirements. Specifically, in Electronic Warfare (EW) applications, the signals usually operate in the GHz range but real-time, high-resolution capability is a necessity [65]. Currently no EW receiver processes signals entirely in digital. The proposed DFT-based MNM with its low computational complexity, is expected to provide the desired high-resolution capability to future digital EW receivers.

2. **Signal-Subspace Information from the Autocorrelation Matrix Only** : The strength of the Minimum-Norm framework as a high-resolution method really comes from its ability to form the 'noise-

subspace' vector **d** by exploiting the orthogonality property in $(II.29)$. It appears that as long as $\mathbf{U}_S$ has some component of the signal-subspace **T**, the solution of $(II.29)$ would retain its high-resolution capability. The DFT-of-AC is an appropriate candidate to produce $\mathbf{U}_S$ because it is a linear combination of the signal-vectors in **T**. This can be seen by rewriting the DFT-of-AC matrix,

$$\mathbf{F} = \mathbf{CD} = \mathbf{T}\left[\frac{1}{M}\sum_{m=1}^{M}\mathbf{a}_m(\mathbf{x}_m^H\mathbf{D})\right]. \tag{$II.36$}$$

In fact, the AC matrix itself is also a possible candidate for obtaining the 'signal-subspace' $\mathbf{U}_S$, because it can be expressed as a linear combination of the signal-vectors in **T**,

$$\mathbf{C} = \mathbf{T}\left[\frac{1}{M}\sum_{m=1}^{M}\mathbf{a}_m\mathbf{x}_m^H\right]. \tag{$II.37$}$$

Not surprisingly, when $\mathbf{U}_S$ is formed with the $p$ largest norm vectors of the estimated **C**, MNM again demonstrated high-resolution capability in simulations (not included). This simpler procedure to obtain 'signal-subspace' information needs to be studied further. But it must be stated that D-MNM performs better at low SNR because the DFT operation accentuates the signal-subspace, as discussed next.

3. **Analysis of the DFT-based Signal Subspace for Arbitrary AOA/Frequency** : For ideal noise-free observations if the frequencies are not on the DFT bins, the DFT-of-AC operation can be expressed as :

$$\begin{aligned}\mathbf{F} &= \mathbf{CD} && (II.38a)\\ &= \mathbf{T\Sigma T}^H\mathbf{D} && (II.38b)\\ &= \mathbf{T\Sigma}\begin{bmatrix}\mathbf{t}_1^H\mathbf{D}\\ \mathbf{t}_2^H\mathbf{D}\\ \vdots\\ \mathbf{t}_p^H\mathbf{D}\end{bmatrix}. && (II.38c)\end{aligned}$$

Consider the matrix at right. Each of the $\mathbf{t}_i^H\mathbf{D}$ vectors are complex valued DFT of a sequence of a complex sinusoid. The magnitude of each row vector, $\mathbf{t}_i^H\mathbf{D}$ has a *Sinc* envelope with a peak occurring at the column corresponding to the bin location closest to the frequency $\omega_i$. For infinite aperture with $N \rightarrow \infty$, *i.e.*, for large number of sensors, each row vector peaks at $\omega_i$ and the other elements of that row approaches zero. The same will be the case for each of the other row vectors also. Hence, asymptotically, the DFT-of-AC operation again produces $p$ largest norm vectors at the true signal frequencies. The asymptotic analysis for the noisy case as defined by (II.25) would also provide similar results. For finite $N$, because of the Sinc weighting, the largest norm vectors will also have contributions from some other $\mathbf{t}_i$ vectors in the **T**. But those components also contain signal-subspace information which is orthogonal to **d** and hence useful for obtaining the minimum-norm vector **d**.

4. **Performance and Accuracy Analysis** : The results presented here indicate that the DFT-of-AC operation retains significant signal information comparable to signal eigenvectors produced by eigendecomposition. This phenomenon needs to be quantified analytically. A possibility would be to analyze and compare the respective Frobenius norms of the Projections onto the true signal basis-space as produced by the signal-subspaces due to the eigen-based as well as DFT-based methods. Most of the existing eigen-based methods have been analyzed to study their performance and accuracy [22, 25, 43, 76, 77]. Following this trend we plan to perform statistical analysis of the bias, variance and the resolution threshold of the estimates produced by the present method.

5. **Estimation of the Parameters of Damped Sinusoids in Noise** : Many eigen-based methods have been successfully utilized in estimating the unknown parameters of damped sinusoids from noisy observations [27, 28]. It appears that with some simple modifications the proposed DFT-based approach could also be used for the same purpose. The advantage would again be that no eigendecomposition but the performance will be comparable.

6. **Largest Norms vs. Peaks** : In all the simulations presented here, the signal subspaces have been formed by selecting the $p$ unit-vectors having largest norms. But the ideal solution may be to pick the unit vectors corresponding to the $p$ largest peaks (having smaller norm vectors on both adjacent bins). This may eliminate any possibility of picking multiple vectors from the vicinity of strong signals. It should be emphasized though that largest norm criteria has worked quite well so far, as demonstrated by a large number of simulations. But this aspect certainly needs further analysis.

7. **Zero padding** : In classical spectral estimation, Periodogram relies on DFT/FFT, but it is often necessary to extend (or, pad) the available data with zeros so that interpolated values between available bins can be calculated. Zero-padding is also used to extend data-lengths to powers of two such that the computational efficiency of the FFT can be taken advantage of. In the simulation studies, no zero-padding had been incorporated so far. It is not quite apparent whether the zero-padding should be done directly to the data or to the covariance estimates and this aspect needs further study. It would also be necessary to study the possible effects on the signal-subspace produced by the DFT-of-AC operation after zero-padding is introduced.

8. **Windowing** : In classical spectral estimation, in order to avoid sudden discontinuities, the observed data is often weighted (or tapered at both ends) by non-rectangular window which tends to enhance the 'dynamic range' at the cost of 'resolution' [19]. In the simulation results presented here, no windowing has been used. But windowing is known to be highly effective in locating weak frequency components which tend to get submerged by the sidelobes of strong components. Though it is believed that that orthogonality property in ($II$.29) is the main contributing factor for the high-resolution capability of D-MNM, it would certainly be interesting to study what effects windowing might have on the performance of D-MNM.

9. **Use of DFT-Based Signal-Subspace in other Eigen-Based Methods** : Other than the Minimum-Norm Method covered in this proposal, there is a large body of work where some form of eigendecomposition is utilized to estimate AOA/Frequencies [1-3, 6, 17, 20-22, 25-29, 33, 39, 41-43, 45, 50, 51, 53, 57, 69, 70, 74-77, 79]. Among the more important results are, MUSIC [50, 51], SVD [28, 29] and ESPRIT [41]. It is quite possible that the proposed DFT-based signal-subspace may be incorporated with some of these existing eigendecomposition based methods, in order to implement those methods without eigendecomposition. Clearly, the proposed approach can be used to implement MUSIC, except that the noise subspace $\mathbf{U}_N$ defined in $(II.26c)$ would have to be utilized. Also, the left and right eigenvectors of the SVD of a data matrix are actually the eigenvectors of correlation matrices. Hence, it appears that some of the SVD-based approaches may also be modified to incorporate DFT-based signal/noise subspaces. Care should be taken about the choice of either the left or right signal-spaces, because both may not contain signal information. The case is not so apparent for those methods which use generalized eigendecomposition [41, 57, 75]. Some of these possibilities need to be further investigated.

10. **Model Order Selection** : In its current form, the proposed approach assumes that the number of targets $(p)$ is known. But Fig. 1 suggests that it may be possible to estimate the model order from the norm of the DFT-of-AC vectors. This possibility needs to be explored further.

11. **DFT-Prony** : There has been some recent interest in implementing the Prony's algorithm in the Frequency-Domain [48]. Clearly, the signal-vectors in $\mathbf{U}_S$ can be treated as multiple time-series to form a $(p+1) \times (p+1)$ covariance matrix (using forward-backward approach) and then the $p$-th order Prony's polynomial can be estimated. Based on preliminary simulations (not included), this approach appears to be simplest of all existing methods with moderately good high-resolution performance. The performance of DFT-Prony is much better than that of the standard Prony's method because the DFT-based signal subspace is cleaned-up though without any eigendecomposition. These ideas needs to be further studied.

12. **Two-Dimensional Frequency-Wavenumber Estimation** : In some array processing scenarios, both the AOAs (related to wavenumbers) and the center frequencies need to be estimated [26, 33, 48-50]. Many existing 1-D eigen-based methods have been extended to 2-D to address this problem. It appears that the DFT-of-AC vectors can be formed in both domains and two $D(z)$ polynomials can be be formed to estimate the the frequencies and AOAs separately. Incorporation of the DFT-based signal-spaces for 2D frequency estimation needs to be further investigated.

13. **Hardware Implementation** : Perhaps the most important and useful practical impact of the proposed method would be in the area of hardware implementation for high-resolution Angles-of-Arrival or frequency estimation. All the currently available methods with good-enough high-resolution capability, rely on some form of iterative optimization or iterative computation of eigenvectors [1-3, 6, 17, 20-22, 25-29, 33, 39, 41-43, 45, 50, 51, 53, 57, 69, 70, 74-77, 79]. In contrast, all that the proposed approach

requires to form the ‘signal-subspace’ is a single matrix multiplication. Furthermore, the matrix to be multiplied is a DFT matrix and it has special structures so that FFT based processing may be utilized to further reduce the computational burden. Hence, one of the major goals in future work would be to devise appropriate strategies to design, develop and, if possible, fabricate VLSI hardware for high-resolution AOA/Frequency estimation.

## III. : Maximum Likelihood Estimation of Multiple Frequencies with Exact Constraints to guarantee Unit Circle Roots

### III.1 : Introduction

Estimating the underlying parameters of multiple complex exponential signals in noise, remains one of the vigorously researched topics in signal processing literature [1-13, 15-35, 37-48, 50-65, 69-79]. For a single sinusoid or when the multiple frequencies are well-separated, the Periodogram performs reasonably well. But if the frequencies are closely spaced, which often occurs when the data length is limited or the aperture is too small, the Periodogram fails to distinguish the frequencies and produces merged frequency estimates. In order to overcome the Periodogram's resolution limitation, many high-resolution methods have been developed in the past two decades [1-13, 15-35, 37-48, 50-65, 69-79]. In contrast to Periodogram, these methods make effective use of some underlying property of the true sinusoidal signal model.

Among all the existing high-resolution frequency estimation methods, the MLE appears to provide the most accurate frequency estimates and has the lowest SNR threshold [5, 31, 32, 46, 47, 55, 61, 78]. Most high-resolution methods rely on the rank and signal or noise subspace information which are extracted from the eigendecomposition of covariance matrix or SVD of data matrix [1-13, 15-35, 37-48, 50-65, 69-79]. On the other hand, the MLE considers the exact model of the exponential signal and attempts to maximize the exact Likelihood function to estimate the unknowns. For a single sinusoid, the peak of the periodogram itself corresponds to the ML estimate, but for multiple exponentials the MLE turns out to be a nonlinear optimization problem [5, 31, 32, 46, 47, 55, 61, 78].

A recently proposed Maximum-Likelihood Estimator (MLE) of multiple exponentials, developed independently in [31] (KiSS) and [5] (IQML), converts the frequency estimation problem into a problem of estimating the coefficients of a $z$-polynomial with roots at the desired frequencies [5, 31]. In polynomial domain, the optimization problem turns out to be *quasi-linear* where a weighted-quadratic criterion is minimized iteratively. Theoretically, the roots of the estimated polynomial should fall right on the unit circle. Though effective to a large extent, KiSS-IQML, as originally proposed, is known to possess one fundamental drawback : the optimization procedure in [5, 31] does not impose sufficient theoretical constraints on the polynomial coefficients for the estimated roots to fall on the unit circle. The primary goal of this part of the work is to address this unresolved problem in KiSS-IQML.

Two conditions must be satisfied for a general $p$-th order $z$-polynomial to have $p$ unit circle roots : conjugate symmetry (C1) and a derivative constraint (C2), the details of which are given later. In KiSS-IQML, only C1 was imposed. The derivative constraint makes the problem highly nonlinear and hence, C2 could not be incorporated in the weighted-quadratic framework of KiSS. But when $p > 1$, C1 alone is not sufficient for unit circle roots. Furthermore, from the theory of Linear-Phase FIR filters, it is well-known that the roots of a symmetric $z$-polynomial may fall either on the unit circle or they may be in reciprocal pairs falling inside and outside of the unit circle. In fact, it was demonstrated in [1] and [3] that, if SNR $\leq$ 10dB and the frequencies are spaced closely, the roots produced by KiSS-IQML were sometimes in reciprocal pairs. In such cases, two frequencies merge to produce only a single frequency estimate. The alternate approach proposed in this paper attempts to alleviate this limitation in KiSS.

There is one particular exception to the two conditions stated above : for $p = 1$, the conjugate symmetry constraint (C1) alone is *sufficient* for the single root to fall on the unit circle. This is the main idea which will be utilized in developing the proposed Constrained-KiSS (C-KiSS) algorithm. Specifically, C1 will be imposed on each of the 1st-order factors of the $p$-th order $z-$polynomial, such that each individual root falls on the unit circle. This process need not be applied to all the frequencies at all SNRs. The constraints are imposed only on those 1st-order factors which produce merged frequency estimates at convergence of KiSS-IQML. The factors for which the roots are already on the unit circle, are held fixed. The proposed algorithm may be considered to be a polynomial-domain counterpart of the 'Alternating Projection' approach [66] where the ML criterion was minimized *w.r.t.* one frequency at a time while the other frequencies were held at the previously estimated values. To the best knowledge of the author, this work appears to be the first attempt to guarantee unit circle roots on the polynomial coefficients for Maximum-Likelihood frequency estimation. The constraints are primarily effective at low SNR levels when there is a higher possibility for KiSS-IQML to produce merged frequency estimates. In simulations, the RMS values of the frequency estimates using C-KiSS were found to be closer to the theoretical CR bounds than those of the original KiSS algorithm.

### III.2 : The Maximum Likelihood Problem and a Brief Overview of KiSS-IQML

The observation samples of a complex multiple exponential signal can be represented as,

$$\mathbf{x}(n) \triangleq \sum_{k=1}^{p} a_k e^{j(\omega_k n \; + \; \phi_k)} + z(n) \quad n = 0, 1, \ldots, N-1 \qquad (III.1)$$

where, $\omega_k$, $a_k$ and $\phi_k$ are the unknown angular frequency, amplitude and phase, respectively, of the $k^{th}$ sinusoid; $p$ is the assumed number of sinusoids and $z(n)$ represents *i.i.d.* $N(0, \sigma^2)$ Gaussian noise samples. For this signal model, the MLE corresponds to optimization of the following error criterion [5, 31, 32, 46, 47, 55, 61, 78].

$$\min_{\omega_1,\ldots,\omega_p,A_1,\ldots,A_p} ||\mathbf{e}||^2 \triangleq \min_{\omega_1,\ldots,\omega_p,A_1,\ldots,A_p} ||\mathbf{x} - \mathbf{Ta}||_2^2 \qquad (III.2)$$

where,

$$\mathbf{x} \triangleq \begin{pmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{pmatrix} \triangleq \mathbf{Ta} \triangleq \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & \cdots & e^{j\omega_p} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\omega_1(N-1)} & e^{j\omega_2(N-1)} & \cdots & e^{j\omega_p(N-1)} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \end{pmatrix}. \qquad (III.3)$$

$A_k \triangleq a_k e^{j\phi_k}$, for $k = 1, 2, \ldots, p$, respectively, are the complex amplitudes. The MLE problem stated in (III.2) is a nonlinear optimization problem with respect to the angular frequencies. Instead, KiSS-IQML forms an alternative but equivalent error criterion in the polynomial coefficient domain which has a quasi-linear structure which is well-suited for iterative optimization. A brief summary of the KiSS-IQML criterion is in order.

Let, $B(z) \triangleq b_0 + b_1 z^{-1} + \ldots + b_p z^{-p}$, be a $p^{th}$ degree $z$-polynomial with $p$ roots at $e^{j\omega_1}$, $e^{j\omega_2} \ldots e^{j\omega_p}$, respectively, and $\mathbf{b} \triangleq [b_0 \; b_1 \; \cdots \; b_p]^T$ be the coefficient vector. The KiSS-IQML criterion for estimating $\mathbf{b}$ is given by,

$$\min_{\{b_i\}_{i=0}^{p}} E(\mathbf{b}) = \mathbf{b}^H \mathbf{X}^H (\mathbf{B}\mathbf{B}^H)^{-1} \mathbf{X}\mathbf{b} \quad \text{where,} \qquad (III.4)$$

$$\mathbf{B} \triangleq \begin{pmatrix} b_p & \ldots & b_1 & b_0 & & 0 \\ & \ddots & & \ddots & \ddots & \\ 0 & & b_p & \ldots & b_1 & b_0 \end{pmatrix}, \qquad \mathbf{X} \triangleq \begin{pmatrix} x(p) & \ldots & x(0) \\ x(p+1) & \ldots & x(1) \\ \vdots & \ddots & \vdots \\ x(N-1) & \ldots & x(N-p-1) \end{pmatrix}. \qquad (III.5)$$

The criterion in (III.4) appears to be quadratic in $\mathbf{b}$, except that the weight matrix itself depends on the unknown coefficients. Hence, this criterion is minimized iteratively. At the $(k-1)$-th iteration :

$$\min_{\mathbf{b}} \; \mathbf{b}^H [\mathbf{X}^H (\mathbf{B}^{(k-1)} \mathbf{B}^{H(k-1)})^{-1} \mathbf{X}] \mathbf{b} \qquad (III.6)$$

is optimized, where the weight matrix $(\mathbf{B}\mathbf{B}^H)$ is formed using the estimate of $\mathbf{b}$ found at the previous iteration. At convergence of these iterations, the frequencies are found from the roots of the estimated polynomial $\hat{B}(z)$. Unfortunately, direct optimization of the criterion in (III.4) does not guarantee that the roots of $\hat{B}(z)$ will indeed fall on the unit circle and it was recognized in [31, 32, 55] that two conditions, as stated next, must be satisfied to guarantee unit circle roots.

### III.3 : Two Conditions for Guaranteeing Unit Circle Roots

**C1** : The coefficients should obey conjugate symmetry constraints, *i.e*,

$$b_k = b^*_{p-k}, \qquad \text{for, } k = 0, 1, \ldots, p, \qquad \text{and,} \qquad (III.7)$$

**C2** : For $p > 1$, the derivative of $B(z)$, *i.e.*,

$$B'(z) \triangleq \frac{\partial B(z)}{\partial z^{-1}} \qquad (III.8)$$

must have zeros either inside or on the unit circle. KiSS-IQML, as originally proposed [5, 31], imposes the conjugate symmetry constraint only. C2 makes the optimization problem highly nonlinear and the weighted-quadratic structure of (III.4) is lost if C2 is incorporated in the algorithm. Hence, no attempt was made in [5, 30-32, 49] to include C2 in the algorithm. But if $p > 1$, C1 is not a sufficient condition for unit circle roots. The same condition may, in fact, lead to roots in reciprocal pairs which can and does occur in KiSS-IQML, especially at low SNR. In such cases, two closely spaced frequencies are estimated as a *single* frequency [31, 32, 55] only.

**III.3.1 : Important Observation** : Interestingly, for $p = 1$, the conjugate symmetry alone is a *sufficient* condition to ensure unit-circle root. Hence, we propose to impose C1 sequentially on each 1st-order factor of $B(z)$ during optimization of (III.4). In that case, the optimization at each step will be with respect to only a 1st-order factor of $B(z)$ and hence, there is no need for satisfying C2.

### III.4 : Constrained KiSS (C-KiSS)

The $p$-th order polynomial $B(z)$ can be expressed in factored form as :

$$B(z) = B^{(p-i)}(z)B^{(i)}(z), \qquad (III.9)$$

where, $B^{(p-i)}(z) \triangleq b_0^{(p-i)} + b_1^{(p-i)}z^{-i} + \ldots + b_{p-1}^{(p-i)}z^{-p+1}$ and $B^{(i)}(z) \triangleq b_0^{(i)} + b_1^{(i)}z^{-1}$, are $(p-1)$-th order and 1st-order factors, respectively. If conjugate symmetry is imposed on the 1st order factor, then, $B^{(i)}(z) = b_0^{(i)} + b_0^{*(i)}z^{-1}$. Note that in (III.9) the coefficients of the polynomial $B(z)$ is formed as the convolution of the coefficients of $B^{(p-i)}(z)$ and $B^{(i)}(z)$. Hence, in matrix-vector notation :

$$\mathbf{b} = \begin{pmatrix} b_0^{(p-i)} & 0 \\ b_1^{(p-i)} & b_0^{(p-i)} \\ \vdots & \vdots \\ b_{p-1}^{(p-i)} & b_{p-2}^{(p-i)} \\ 0 & b_{p-1}^{(p-i)} \end{pmatrix} \begin{pmatrix} b_0^{(i)} \\ b_0^{*(i)} \end{pmatrix} \triangleq \mathbf{B}_{p-i} \begin{pmatrix} 1 & j \\ 1 & -j \end{pmatrix} \begin{pmatrix} b_{0r}^{(i)} \\ b_{0i}^{(i)} \end{pmatrix} \triangleq \triangleq \mathbf{B}_{p-i}\mathbf{J}\mathbf{b}_i, \qquad (III.10)$$

where, $\mathbf{B}_{p-i}$ denotes the matrix-factor with the $i$-th 1-st order factor removed and $b_0^{(i)} \triangleq b_{0r}^{(i)} + jb_{0i}^{(i)}$. Using (III.10) in (III.6), each 1st-order factor of $B(z)$ is estimated at the $k$-th iteration by optimizing,

$$\min_{\mathbf{b}_i} \; \mathbf{b}_i[\mathbf{J}^H\mathbf{B}_{p-i}^{H\;(k-1)}\mathbf{X}^H(\mathbf{B}^{(k-1)}\mathbf{B}^{H(k-1)})^{-1}\mathbf{X}\mathbf{B}_{p-i}^{(k-1)}\mathbf{J}]\mathbf{b}_i, \qquad \text{for, } i = 1, 2, \ldots, p. \qquad (III.11)$$

This is a weighted-quadratic criterion of the form :

$$\mathbf{b}_i^H\mathbf{W}_{p-i}^{(k-1)}\mathbf{b}_i \qquad \text{where,} \qquad (III.12a)$$

$$\mathbf{W}_{p-i}^{(k-1)} \triangleq \mathbf{J}^H\mathbf{B}_{p-i}^{H\;(k-1)}\mathbf{X}^H(\mathbf{B}^{(k-1)}\mathbf{B}^{H(k-1)})^{-1}\mathbf{X}\mathbf{B}_{p-i}^{(k-1)}\mathbf{J} \qquad (III.12b)$$

is the weight matrix formed with the estimates found at the previous iteration step. The criterion in (III.11) can be optimized sequentially or concurrently for each $i$-th first order factor. At each iteration, $\mathbf{b}_i$ is estimated as the eigenvector corresponding to the minimum eigenvalue of $\mathbf{W}_{p-i}^{(k-1)} \in \mathbb{R}^{2\times 2}$. The advantage of using (III.12a) instead of (III.6) is that, since each $B^{(i)}(z)$ is a first-order $z$-polynomial, *only* the conjugate symmetry constraint is sufficient to guarantee the root of $B^{(i)}(z)$ to fall on the unit circle. In practice, the alternate optimization procedure in (III.11) need not be carried out for all the $p$ factors of $B(z)$. It needs to be invoked only in those cases for which KiSS-IQML produces merged frequency estimates. The roots which are already on the unit circle need not be optimized further. This sequential process guarantees that all the roots of $B(z)$ will indeed fall on the unit circle while the exact ML criterion is also optimized at the same time.

### III.5 : Simulation Results

The algorithm described above has been tested with the same simulated data set used in [28, 29, 31, 69]. The following formula was used to generate the data,

$$x(n) = a_1 e^{j\omega_1 n} + a_2 e^{j\omega_2 n} + z(n) \qquad (III.13)$$
$$n = 0,\ 1,\ \ldots, 24$$

where, $\omega_1 = 2\pi f_1$, $\omega_2 = 2\pi f_2$, $f_1$ and $f_2$ being 0.52 and 0.50, respectively, $A_1 = 1$, $A_2 = e^{j\frac{\pi}{4}}$, $z(n)$ is a computer generated white zero-mean, complex gaussianly distributed noise sequence with variance $= \sigma^2$, *i.e.*, $\frac{\sigma^2}{2}$ is the variance of the real and the imaginary parts of $z(n)$. SNR is defined as, 10 $\log_{10}(\frac{|a_i|^2}{\sigma^2})$. Two hundred data sets with independent noise epochs were used.

Fig. 6a and 6b show the estimated roots for 200 independent trials of KiSS-IQML for SNR = 5dB and 10dB, respectively. Fig. 6d and 6e show the corresponding results with C-KiSS. For the 10dB case, Figures 6c and 6f show only the merged cases before after applying the exact constraints. The unit circle roots in Fig. 6f does show wider spread than the corresponding merged frequency estimates in Fig. 6c. Fig. 7 compares the performance of KiSS-IQML and C-KiSS with the theoretical CR bound. The results verify that C-KiSS performs better than original KiSS at low SNR range.

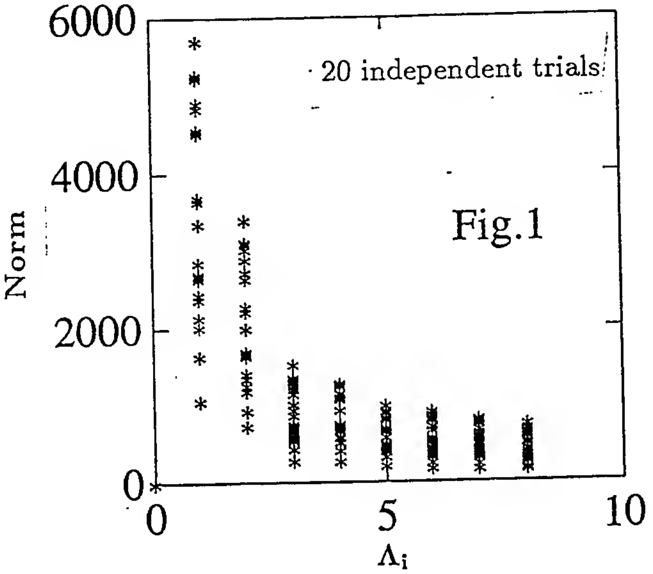


Fig.1. Norms of the DFT-of-AC vectors

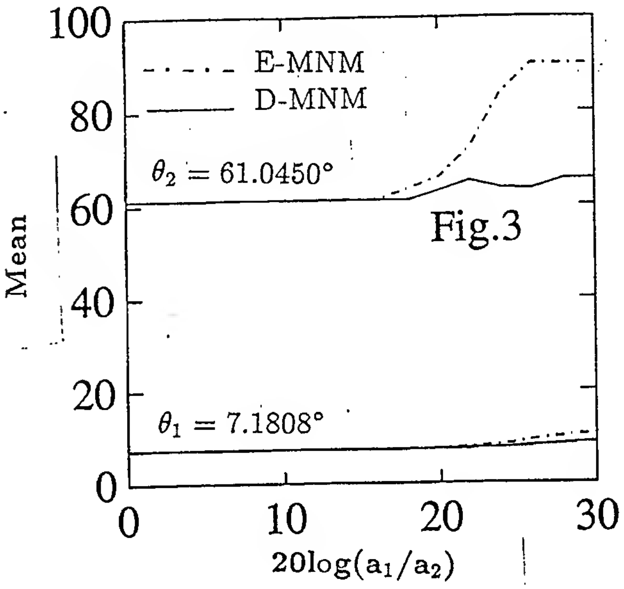


Fig.3. Means of $\theta_1$ and $\theta_2$ for 50 independent trials

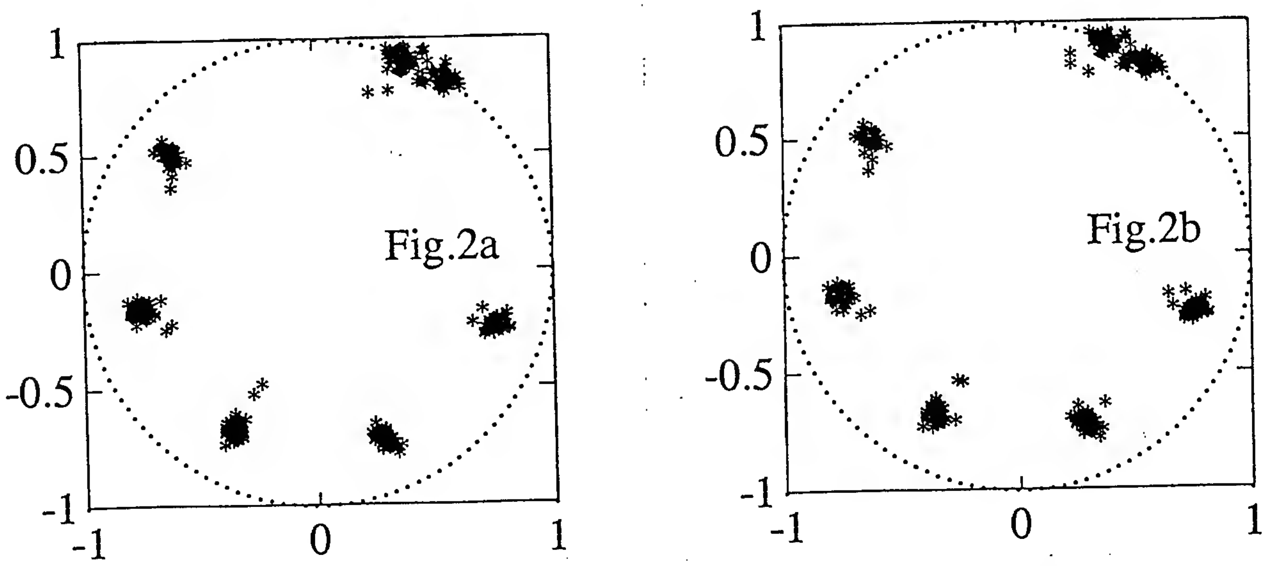


Fig.2. Roots of $D(z)$ using (a) E-MNM and (b) D-MNM for 50 independent.

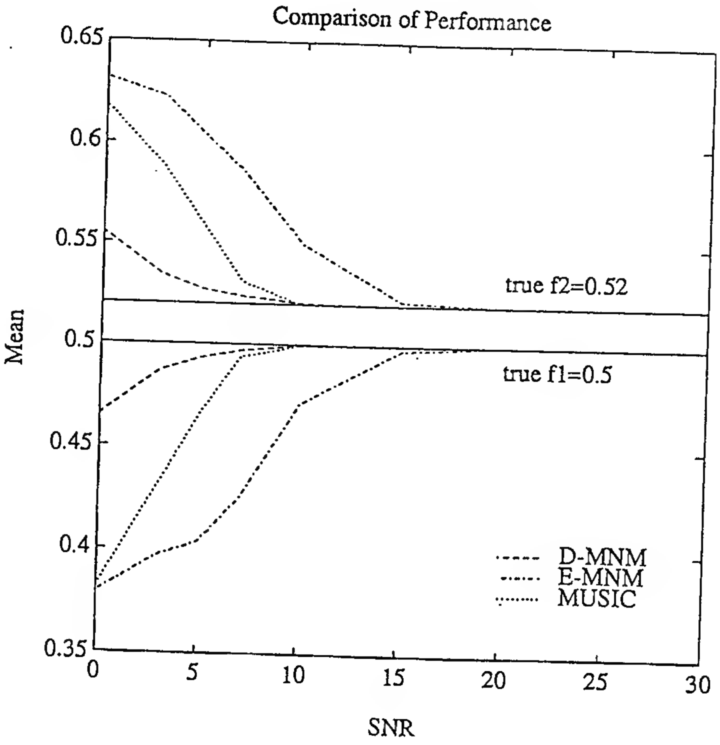


Fig.4. Comparison of Mean values with 500 independent trials for three methods.

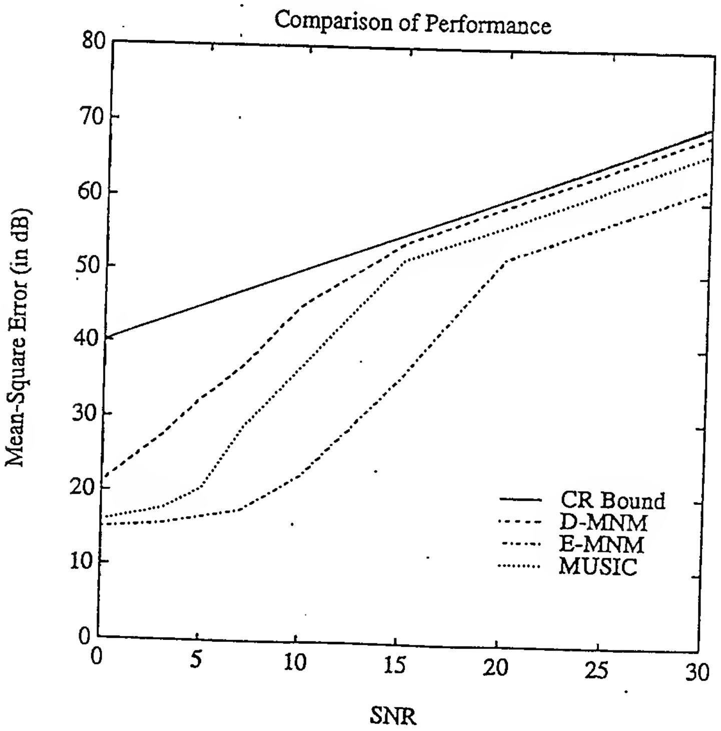


Fig.5. Comparison of RMS values with CR bounds for 500 independent trials.

| SNR | Successful Trials | | Bias (in degrees) | | RMS | |
|---|---|---|---|---|---|---|
| (in dB) | D-MNM | E-MNM | D-MNM | E-MNM | D-MNM | E-MNM |
| 5 | 59 | 39 | -0.8480 | -0.5539 | 1.4311 | 1.3623 |
| | | | 1.1589 | 0.4329 | 1.9174 | 1.9322 |
| 10 | 139 | 130 | -0.3154 | -0.4589 | 1.3529 | 1.5063 |
| | | | 0.8940 | 0.7603 | 1.7910 | 1.8571 |
| 15 | 191 | 189 | -0.0714 | -0.1094 | 0.9812 | 1.0021 |
| | | | 0.4623 | 0.3648 | 1.3118 | 1.3212 |
| 20 | 199 | 198 | -0.0055 | -0.0252 | 0.6777 | 0.6822 |
| | | | 0.1717 | 0.1170 | 0.8440 | 0.8017 |
| 25 | 200 | 200 | 4.99e-4 | -0.0067 | 0.4129 | 0.4302 |
| | | | 0.0611 | 0.0481 | 0.4820 | 0.4826 |
| 30 | 200 | 200 | 0.0037 | 0.0018 | 0.2297 | 0.2329 |
| | | | 0.0263 | 0.0219 | 0.2728 | 0.2737 |

Table 1 : Comparison of performance of D-MNM and E-MNM.

| SNR | Bias (in degrees) | | | RMS | | |
|---|---|---|---|---|---|---|
| (in dB) | D-MNM | E-MNM | MUSIC | D-MNM | E-MNM | MUSIC |
| 0 | -0.0349 | -0.1205 | -0.1178 | 0.0876 | 0.1783 | 0.1594 |
| | 0.0352 | 0.1118 | 0.0983 | 0.0786 | 0.1748 | 0.1486 |
| 3 | -0.0133 | -0.1029 | -0.0681 | 0.0415 | 0.1654 | 0.1312 |
| | 0.0141 | 0.1027 | 0.0678 | 0.0468 | 0.1640 | 0.1265 |
| 5 | -0.0070 | -0.0964 | -0.0343 | 0.0232 | 0.1476 | 0.0946 |
| | 0.0072 | 0.0838 | 0.0392 | 0.0342 | 0.1378 | 0.0991 |
| 7 | -0.0031 | -0.0754 | -0.0063 | 0.0142 | 0.1322 | 0.0373 |
| | 0.0039 | 0.0658 | 0.0111 | 0.0245 | 0.1189 | 0.0560 |
| 10 | -3.40e-4 | -0.0289 | -5.62e-4 | 0.0054 | 0.0756 | 0.0140 |
| | 6.54e-4 | 0.0301 | -1.19e-4 | 0.0093 | 0.0776 | 0.0058 |
| 15 | -7.10e-5 | -0.0023 | 2.80e-5 | 0.0020 | 0.0159 | 0.0026 |
| | -1.82e-4 | 0.0019 | -9.05e-5 | 0.0022 | 0.0134 | 0.0026 |
| 20 | -3.40e-6 | -1.04e-5 | 1.61e-5 | 0.0011 | 0.0025 | 0.0015 |
| | -7.76e-5 | 6.34e-5 | -5.23e-5 | 0.0012 | 0.0025 | 0.0014 |
| 30 | 8.64e-6 | 2.18e-5 | 3.35e-6 | 3.53e-4 | 7.87e-4 | 4.61e-4 |
| | -1.77e-5 | -9.01e-6 | -1.51e-5 | 3.75e-4 | 7.84e-4 | 4.50e-4 |

Table 2. Comparison of Bias and RMS values for three methods with 500 independent trials.

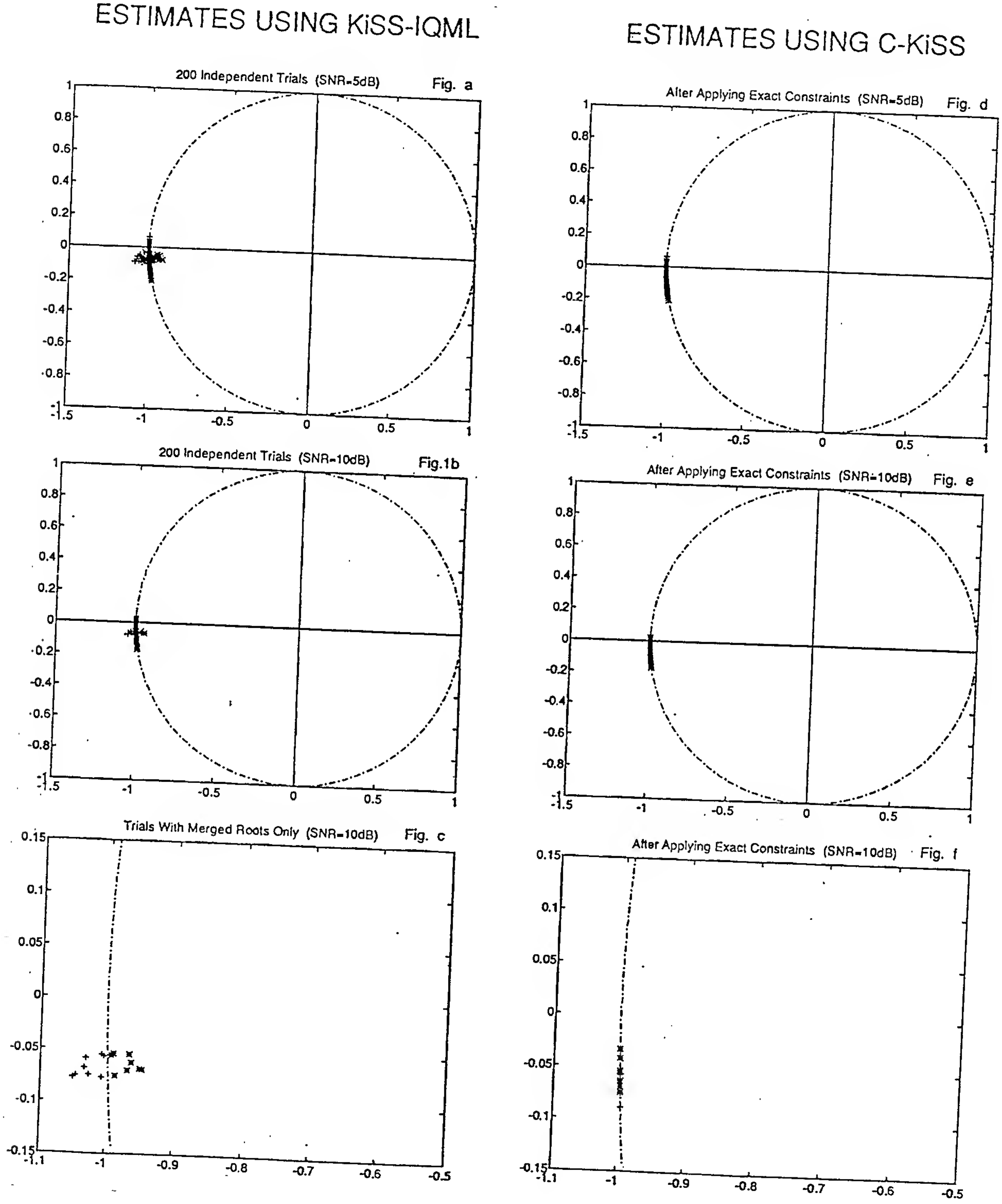


Fig. 6 Superimposed plots of estimated roots for 200 independent trials using KiSS-IQML (a-c) and C-KiSS (d-f).

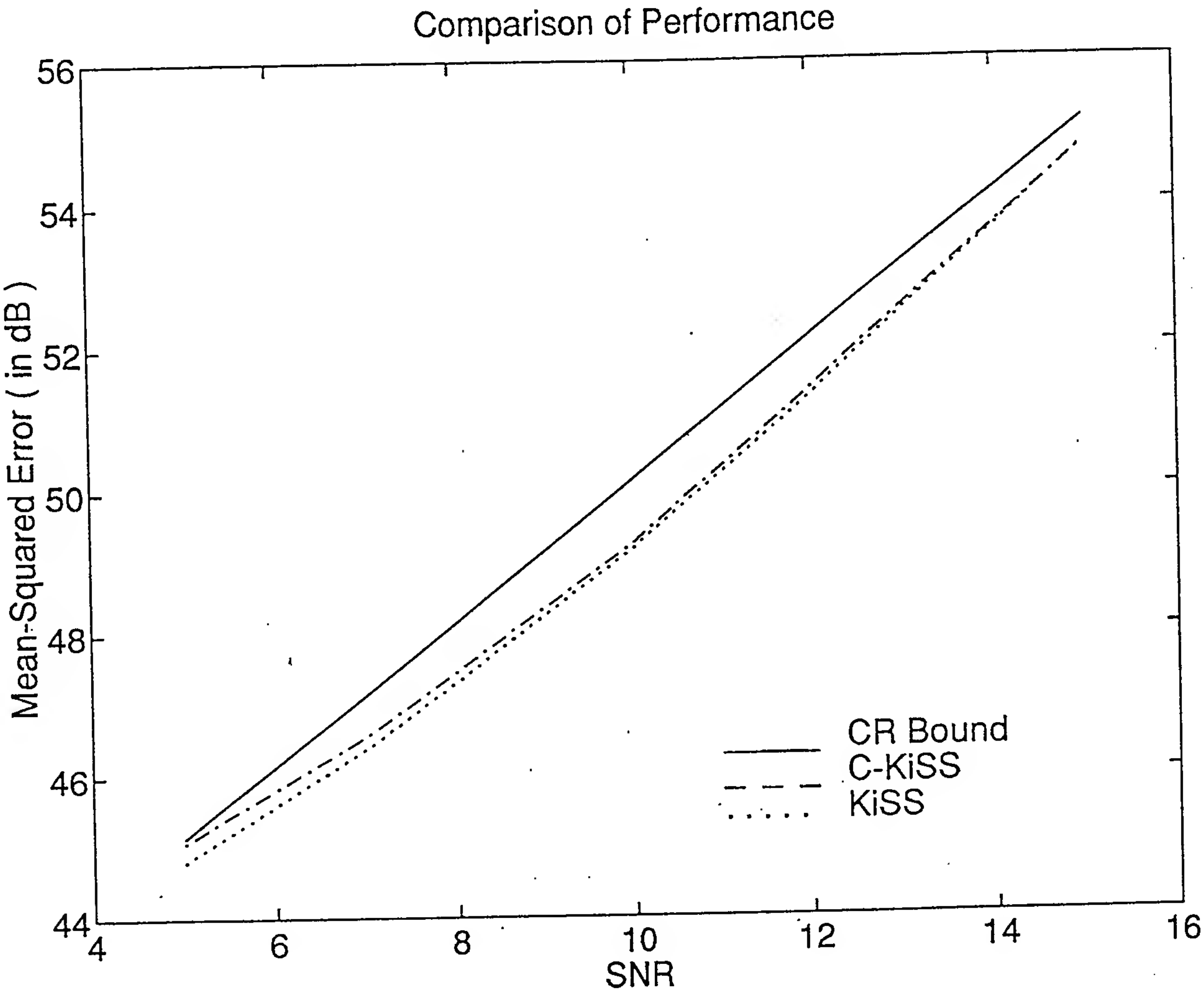


Fig. 7 Performance comparison of KiSS-IQML and C-KiSS with the theoretical CR-bound. 200 independent trials were used.

## References

[1] G. Bienvenue and L. Kopp, "Principle de la goniometric od passive adaptive," *Proceedings 7'eme Colloque GRESTI*, Nice, France, pp. 106/1-106/10, 1979.

[2] G. Bienvenue and L. Kopp, "Adaptivity to Background Noise Spatial Coherence for High Resolution Passive Methods," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Denver, Colorado, pp. 307-310, 1980.

[3] G. Bienvenue and L. Kopp, "Decreasing High Resolution Method Sensitivity by Conventional Beamformer Preprocessing," *Proceedings of International Conference of Acoustics, Speech and Signal Processing*, pp. 33.2.1-4, 1984.

[4] R. N. Bracewell, "Radio Interferometry of Discrete Sources," *Proceedings of IRE*, vol. 46, pp. 97-105, Jan. 1958.

[5] Y. Bressler and A. Macovski, "Exact Maximum Likelihood Parameter Estimation of Superimposed Exponential Signals in Noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 10, pp. 1081-1089, Oct., 1986.

[6] K. Buckley and X.-L. Xu, "Spatial Spectrum Estimation in a Location Sector," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-38, no. 11, pp. 1842-1852, Nov., 1990.

[7] J. P. Burg, "Maximum Entropy Spectral Analysis," presented at the *37th Annual International SEG Meeting*, Oklahoma City, OK, 1967.

[8] J. P. Burg, *Maximum Entropy Spectral Analysis*, Ph.D. Dissertation, Stanford University, Stanford, CA, May, 1975.

[9] J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408-1418, 1969.

[10] J. Y. Cheung, "A Direct Adaptive Frequency Estimation Technique," *30th Midwest Symposium on Circuits and Systems*, New York, Aug., 1987.

[11] M. P. Clark and L. L. Scharf, "Reducing the Complexity of Parametric Estimators for Deterministic Modal Analysis," *IEEE Transactions on Signal Processing*. vol. 40, no. 7, pp. 1811-1813, July, 1992.

[12] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Fourier Series," *Math. Comput.*, vol. 19, pp. 297-301, 1965.

[13] S. Haykin et al, Editors, *Array Signal Processing*, Prentice-Hall, 1985.

[14] T. L. Henderson, "Matrix Methods for Determining System Poles from Transient Response," Technical Report, University of Kentucky, Louisville, Kentucky, May, 1980.

[15] T. L. Henderson, "Rank Reduction for Broadband Waves Incident on a Linear Receiving Aperture," *19th Asilomar Conference on Circuits, Systems and Computers*, Nov., 1985.

[16] F. B. Hildebrand, *Introduction to Numerical Analysis*, McGraw-Hill, New York, Chapter 9, 1956.

[17] Y. H. Hu, "Adaptive methods for Real Time Pisarenko Spectrum Estimate," *Proceedings of the ICASSP-1985*, March, 1985.

[18] L. B. Jackson et al., "Frequency Estimation by Linear Prediction," in the *Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing-1979*, Washington, DC, pp. 352-356, Apr., 1979.

[19] L.B. Jackson, *Digital Filters and Signal Processing*, Kluwer, Boston, 1986.

[20] D. H. Johnson et al, "Improving the Resolution of Bearing in Passive Sonar Arrays by Eigenvalue Analysis," Technical Report EE-8102, Department of Electrical Engineering, Rice University, Houston, Texas, 1980.

[21] D. H. Johnson, "The Application of Spectral Methods to Bearing Estimation Problems," *Proceedings of the IEEE*, Vol. 70, pp. 1018-1028, Sept., 1982.

[22] M. Kaveh and A. J. Barbell, "The Statistical Performance of the MUSIC and the Minimum-Norm Algorithms in Resolving Plane Waves in Noise," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-34, pp. 331-341, April, 1986.

[23] S. M. Kay and A. K. Shaw, "Frequency Estimation by Principal Component Autoregressive Spectral Estimator Without Eigendecomposition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1988.

[24] W. C. Knight, R. G. Pridham and S. M. Kay, "Digital Signal Processing of Sonar," *Proceedings of the IEEE*, vol. 69, no. 11, Nov. 1981.

[25] A. C. Kot, S. Parthasarathy, D. W. Tufts and R. J. Vaccaro, "Statistical Performance of Single Sinusoid Frequency Estimation in White Noise Using State-Variable Balancing and Linear Prediction," to be published in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1988.

[26] R. Kumaresan and D. W. Tufts, "A Two-Dimensional Technique for Frequency-Wavenumber Estimation," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1515-1517, Nov., 1981.

[27] R. Kumaresan and D. W. Tufts, "Estimating the Parameters of Exponentially Damped Sinusoids and Pole-Zero Modeling in Noise," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.ASSP-30, no. 6, pp. 833-840, Dec., 1982.

[28] R. Kumaresan, *Estimating the Parameters of Exponentially Damped and Undamped Sinusoidal Signals*, Ph. D. Dissertation, University of Rhode Island, 1982.

[29] R. Kumaresan and D. W. Tufts, "Estimating the Angles of Arrival of Multiple Planewaves," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-19, no .1, pp. 134-139, Jan., 1983.

[30] R. Kumaresan and A. K. Shaw, "High Resolution Bearing Estimation Without Eigendecomposition," *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Florida, April, 1985.

[31] R. Kumaresan, L. L. Scharf and A. K. Shaw, "An Algorithm for Pole-Zero Modeling and Spectral Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.ASSP-34, pp. 637-640, June, 1986.

[32] R. Kumaresan and A.K. Shaw, "Superresolution by Structured Matrix Approximation", *IEEE Transactions on Antennas and Propagation*, Vol. AP-36, pp. 34-44, 1988.

[33] S. Y. Kung, K. S. Arun and D. V. Bhaskar Rao, "State-Space and Singular-Value Decomposition-Based Approximation Methods for the Harmonic Retrieval Problem," *Journal of the Optical Society of America*, vol. 73, pp. 1799-1811, Dec., 1983.

[34] R. T. Lacoss, "Data Adaptive Spectral Analysis Methods," *Geophysics*, vol. 36, pp. 661-675, Aug., 1971.

[35] J. Makhoul, "Linear Prediction : A Tutorial Review," *Proceedings of the IEEE*, vol. 63, pp. 561-580, April, 1975.

[36] M. Marden, *Geometry of Polynomials*, Math Surveys, No. 3, American Mathematical Society, Provi-

dence, RI, pp. 206, 1966.

[37] W. S. McCormick, "A High-Resolution, Near Real-Time Frequency Estimator for Sub-Microsecond Pulses," *IEEE International Conference on Systems Engineering*, Dayton, OH, pp.33-38, Aug., 1989.

[38] A. H. Nuttal, "Spectral Analysis of Univariate Process with Bad Data Points, via Maximum Entropy and Linear Predictive Techniques," NUSC Technical Report, TR-5303, Naval Underwater Systems Center, New London, CT, Mar., 1976.

[39] N. L. Owsley, "Adaptive Data Orthogonalization," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 1978*, pp. 109-112, 1978.

[40] L. C. Palmer, "Coarse Frequency Estimation Using the Discrete Fourier Transform," *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 104-109, Jan., 1974.

[41] A. Paulraj, R. Roy and T. Kailath, "Estimation of Signal Parameters via Rotational Invariance Techniques - ESPRIT," *Nineteenth ASILOMAR Conference on Signals, Systems and Computers*, Pacific Grove, CA, Oct., 1985. Also presented at, *International Conference on Acoustic, Speech and Signal Processing*, pp. 83-89, 1986.

[42] V. F. Pisarenko, "The Retrieval of Harmonics from Covariance Functions," *Geophysical Journal of the Royal Astronomical Society*, Vol. 33, pp. 347-366, 1973.

[43] B. Porat and B. Friedlander [1987], "On the Accuracy of the Kumaresan-Tufts Method for Estimating Complex Damped Exponentials," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-35, no. 2, Feb..

[44] R. Prony, "Essai Experimental et Analytique etc.," L'Polytechnique, Paris, 1 Cahier 2, pp. 24-76, 1795.

[45] S. S. Reddi, "Multiple Source Location- A Digital Approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-15, no.1, pp. 95-105, 1979.

[46] D. C. Rife and R. R. Boorstyn, "Single Tone Parameter Estimation from Discrete-Time Observations," *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 591-598, Sept., 1974.

[47] D. C. Rife and R. R. Boorstyn, "Multiple Tone Parameter Estimation from Discrete Time Observations," *Bell Systems Technical Journal*, vol. 55, pp. 1389-1410, 1976.

[48] L. L. Scharf, *Statistical Signal Processing - Detection, Estimation and Time Series Analysis*, Addison-Wesley, Reading, MA, 1990.

[49] D. Curtis Schleher, *Introduction to Electronic Warfare*, Artech House, MA, 1986.

[50] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *Proceedings of RADC Spectral Estimation Workshop*, pp. 243-258, Rome, New York, 1979.

[51] R. O. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*, Ph. D. Dissertation, Dept. of Elect. Engg., Stanford University, 1981.

[52] A. Schuster, "On the Investigation of Hidden Periodicities with Application to a Supposed 26-Day Period of meteorological Phenomena," *Terr. Magnet.*, vol. 3, pp. 13-41, 1898.

[53] T. J. Shan , M. Wax and T. Kailath, "Spatial Smoothing Approach to Location Estimation of Coherent Sources," *Asilomar Conference on Circuits and Systems and Computers*, Pacific Grove, California, pp. 367-371, Nov., 1983.

[54] A. K. Shaw and R. Kumaresan, "Frequency-Wavenumber Estimation by Structured Matrix Approximation," *Third IEEE-ASSP Workshop on Spectrum Estimation and Modeling*, Boston, MA, pp. 81-84, Nov., 1986.

[55] A.K. Shaw, *Structured Matrix Problems in Signal Processing*, Ph.D. Dissertation, Univ. of Rhode Island, RI, 1987.

[56] A.K. Shaw and R. Kumaresan, "Some Structured Matrix Approximation Problems", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, pp. 2324-2327, April, 1988.

[57] A. K. Shaw and R. Kumaresan, "Estimation of Angles of Arrivals of Broadband Sources," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, pp. 2296-2299, April, 1987.

[58] A. K. Shaw, "A Novel Cyclic Algorithm for Maximum-Likelihood Frequency Estimation," *IEEE International Conference on Systems Engineering*, Dayton, OH, Aug., 1991.

[59] A. K. Shaw, *New Algorithms for Broad-band and Narrow-band Source Localization and A Separable 2-D IIR Filter Realization*, Technical Report, AFOSR-Bolling AFB, Sept., 1991.

[60] A. K. Shaw and S. Nunes, "Detection and Adaptive Frequency Estimation for Digital Microwave Receivers," Final Report, AFOSR Summer Research Program, Oct., 1992.

[61] A. K. Shaw, "Approximate Maximum Likelihood Estimation of Multiple Frequencies with Constraints to Guarantee Unit Circle Roots," *IEEE Transactions on Signal Processing,* to be published, 1994.

[62] A. K. Shaw and W. Xia, "High-Resolution Angles of Arrival Estimation using Minimum-Norm Method Without Eigendecomposition," *IEEE International Conference on Acoustics, Speech and Signal Processing,* Adelaide, Australia, April, 1994.

[63] A. K. Shaw and W. Xia, "Minimum-Norm Method Without Eigendecomposition," *IEEE Signal Processing Letters*, in press, 1994.

[64] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing,* Prentice-Hall, NJ, 1992.

[65] J. B. Y. Tsui, *Digital Microwave Receivers : Theory and Applications*, Artech House, MA, 1989.

[66] J. B. Y. Tsui, *Microwave receivers with Related Components*, National Technical Information Center, 1982, Peninsula, Los Altos, CA, 1985.

[67] J. B. Y. Tsui, *Microwave Receivers with Electronic Warfare Applications*, John Wiley and Sons., New York, 1986.

[68] J. B. Y. Tsui and D. Sharpin, Unpublished Report on Time-Domain Detection for Digital Receivers, June, 1992.

[69] D. W. Tufts and R. Kumaresan, "Frequency Estimation of Multiple Sinusoids : Making Linear Prediction Perform Like Maximum Likelihood," *Proceedings of the IEEE*, vol. 70, pp. 975-989. Sept., 1982.

[70] D. W. Tufts and C. D. Melissinos, "Simple, Effective Computation of Principal Eigenvectors and their Eigenvalues and Application to High-Resolution Estimation of Frequencies," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 10, pp. 1046-1053, Oct., 1986.

[71] D. W. Tufts and S. Parthasarathy, "Statistical Analysis of the Effects of Matrix Perturbation in Some Least Squares Problems," *2nd SIAM Conference on Applied Linear Algebra*, Raleigh, NC, 1985.

[72] T. J. Ulrych and T. N. Bishop, "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," *Rev. Geophysics and Space Physics*, vol. 13, pp. 183-200, Feb., 1975.

[73] T. J. Ulrych and R. W. Clayton, "Time Series Modeling and Maximum Entropy," *Phys. Earth Planetary Int.*, vol. 12, 1976.

[74] R. J. Vaccaro, "On Adaptive Implementations of Pisarenko's Harmonic Retrieval Method," *Proceedings of the ICASSP-84*, March, 1984.

[75] H. Wang and M. Kaveh, "Estimation of Angles of Arrival for Wideband Sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing-1984*, San Diego, California, pp. 7.5.1-7.5.4, Mar. 19-21, 1984.

[76] H. Wang and M. Kaveh, "On the Performance of Signal Subspace Processing-Part I: Narrowband Systems," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 5, pp. 1201-1209, Oct., 1986.

[77] X.-L. Xu and K. M. Buckley, "Bias Analysis of the MUSIC Location Estimator," *IEEE Transactions on Signal Processing.* vol. 40, no. 10, Oct., 1992.

[78] I. Ziskind and M. Wax, "Maximum Likelihood Localization of Multiple Sources by Alternating Projection," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-36, no. 10, pp. 1553-1560, Oct., 1988.

[79] M. D. Zoltowski, G. M. Kautz and S. D. Silverstein, "Beamspace Root-MUSIC," *IEEE Transactions on Signal Processing.* vol. 41, no. 1, Jan., 1993.

# CONTACT LAW AND NUMERICAL MODELING FOR LOW-VELOCITY IMPACT OF COMPOSITE MATERIALS

Rob Slater
Graduate Research Assistant
Department of Mechanical, Industrial, and Nuclear Engineering

University of Cincinnati
Cincinnati, OH 45221-0072

Final Report for:
Summer Research Extension Program
Wright Laboratory

Sponsored by:
Air Force Office Of Scientific Research
Bolling Air Force Base, Washington, D.C.

and
University of Cincinnati
Cincinnati Ohio

December 1993

# CONTACT LAW AND NUMERICAL MODELING FOR LOW-VELOCITY IMPACT OF COMPOSITE MATERIALS

Rob Slater
Department of Mechanical, Industrial, and Nuclear Engineering
University of Cincinnati

## ABSTRACT

Low-velocity impact damage is a critical consideration in the design of composite laminates for aircraft structures. Composite materials offer high stiffness and strength at a significant weight savings over metals. Low-velocity impact damage is a particularly insidious problem because it is difficult to detect by visual inspection. These impacts occur during the course of normal flight and maintenance operations and often leave only a small, shallow dent on the impact surface. But there may be significant interior and backface damage to the laminate. In order to design composite materials for impact tolerance, it is necessary to develop methods to model composite structures and simulate the loads. A contact law describes the relationship between the indentation of an impacting object into the target and the transmitted force. A modified Hertzian theory is commonly used. This model requires experimental data and only predicts the total force as a function of indentation. For more accurate determination of the stresses which cause impact damage, a contact law which predicts the distribution of force in the contact area is necessary. Numerical techniques such as finite elements are efficient methods for performing analyses of low-velocity impact on composite structures. As computer technology continues to evolve, it is necessary to develop and evaluate new techniques for solving these problems. Explicit integration schemes are now becoming widely available. They are recognized for their superiority in dynamics problems involving explosions, very large deflections, and high-speed collisions. They also handle contact problems efficiently and accurately, so investigation of their usefulness on low-velocity impact problems is essential.

# CONTACT LAW AND NUMERICAL MODELING FOR LOW-VELOCITY MODELING OF COMPOSITE MATERIALS

ROB SLATER

## INTRODUCTION

Composite materials are frequently considered in the design of aircraft structures. They offer high stiffness and strength at a significant weight savings over metals. There are distinct differences in the design criteria for metal and composite structures. Fracture and fatigue are the key concerns in the service life of metal structures. Design of composites is driven by concerns such as delaminations at the discontinuities, voids, wrinkles, and low-velocity impact by foreign objects. Laminated composites are particularly susceptible to low-velocity impact damage. These impacts occur during normal operations due to hail or stones blown around a runway by jet engines or during maintenance operations due to tool drop of footsteps. The damage mechanisms associated with low-velocity impacts include matrix cracking, ply delamination, and fiber breakage. Delamination is particularly troublesome because serious internal and back-face damage may be present even though the damage on the impact face appears to be quite minor. It is often not detected during routine visual inspection because the visible damage is slight. Significant reductions in strength and stiffness may result. The design criteria for composites generally require tolerance of a certain level of impact, defined either as an impact energy or a maximum surface indentation, which must be tolerated without significant degradation of service life. The simplest measure of impact is the kinetic energy of the impactor; i.e. an object of known mass impacting at a known velocity. Establishing rational design criteria for low-velocity impacts is the subject of numerous investigations. Characterization of initial damage and its propensity to propagate under service loading are important considerations in the development of the design criteria. In addition, impact may occur at any point of a structure, so a panel must be designed to resist impact at any location. There are numerous local stress raisers in actual aircraft panels such as cutouts, ply drops, and stiffeners which must be considered in designing for impact damage tolerance. It is prohibitively expensive to perform laboratory tests on every composite panel configuration in an aircraft and difficult to correlate results from coupon tests to built-up structures. A more prudent course is to develop analytical criteria and validate them with experiments.

In order to design for impact tolerance, it is necessary to develop methods to model composite structures and simulate the loads. Numerous researchers have attempted to determine a methodology for predicting damage in a composite laminate subject to impact [1-5]. There are a number of variables which must be considered in the analysis of such events. Sensitivity to changes in materials, stacking sequence and panel thickness are significant. Boundary conditions in numerical models and experimental setups are difficult to match except by trial and error. Various failure theories have been offered to predict the size, shape, and location of damage for a given state of stress, but there is little consensus about what mechanisms are at work. In short, no general method exists to predict damage in an arbitrary composite laminate due to an arbitrary impact.

The objective of this research is to develop a simple analytical method for determining the relationship between distribution of contact pressure and indentation of a composite laminate struck by a spherical impactor. Along with such a contact law, methods of modeling the behavior of composite laminates under impact loading, using finite elements and other numerical analysis tools are to be investigated.

## SUMMARY OF RESEARCH

The research conducted this summer was under the supervision of Dr. Ronald Huston, Professor of Mechanics at the University of Cincinnati. Dr. V.B. Venkayya, WL/FIBR, was the focal point for the research at Wright-Patterson Air Force Base. Dr. Greg Schoeppner, WL/FIBC, was also a major contributor of technical advice and direction. The work explored several issues concerning low-velocity impact of composites. First, a comprehensive review of existing literature was performed, focusing on the following topics:

1) Analytic Models
2) Numerical Models
3) Contact Law
4) Damage Prediction
5) Residual Strength

Analytic models concern mainly plate theories which are suitable for modeling laminated composites. It is generally accepted that transverse shear stresses are a key concern for impact problems. The simpler plate theories do not accurately predict these stresses and thus have limited usefulness in solving impact problems. It is normally necessary to use more complex theories. Noor and Burton published a survey article "Assessment of Shear Deformation Theories for Multilayered Composite Plates" [6] which lists and compares numerous methods. Several of the plate theories were reviewed more closely in order to determine which are most appropriate for solution of low-velocity impact problems.

A suitable model for the laminate must be determined. The complexity of this model depends strongly on the type of information desired. If the gross response of the panel is all that is required, a smeared plate theory may be sufficient. But if damage is to be predicted, a discrete laminate theory is necessary to evaluate the stresses layer-by-layer through the thickness. The commonly used first-order shear deformation plate theories lead to discontinuities in the stresses at the interfaces, so a more computationally intense formulation is required.

Finite element methods were the focus of research on numerical methods. Some simple analyses were performed using NASTRAN at Wright-Patterson AFB in 1992. These demonstrated the general feasibility of performing impact analyses using general-purpose finite element programs but revealed that modeling techniques would require further development to provide highly accurate simulations. The ANSYS finite element package has better non-linear capabilities than NASTRAN and includes some elements specifically designed for contact problems, including non-linear springs, gap elements, and contact surfaces.

Finite element methods have frequently been used to evaluate the stress field in an impacted laminate and correlate levels of stress components to experimentally observed damage [7-9]. The stress gradients are very high in the region where the impactor strikes the target and the immediate vicinity, so in order to find the stresses accurately, a refined element mesh is required. The contact force should be distributed over the numerous nodes of this region. As with all analyses involving contact, problems arise in predicting the size of the contact area because it varies non-linearly with transmitted force.

The two codes previously mentioned use implicit schemes to integrate the equations of motion for dynamic analyses. These methods, such as the Newmark method, are unconditionally stable so the user may select the time step according to the desired accuracy of the solution. Stiffness and mass matrices must be assembled and a system of simultaneous equations are solved for each time step. For models with a moderate number of degrees of freedom or solutions in which a very small time step is not required, these methods work well. But for large, complex models and high-speed dynamics problems, the solutions become expensive due to issues of temporary storage for the large matrices and the CPU time needed to perform the matrix operations.

For such problems, explicit integration schemes present an alternative. These methods, including the Central Difference Method, are very efficient for analyses involving high-speed dynamics, contact and separation of bodies or surfaces, and large deformations of impacting bodies. Explicit methods have a minimum time step associated with them to insure stability. It is often one or two orders of magnitude smaller than might be chosen for an implicit scheme, but explicit methods do not require assembly and inversion for each time step. Lumped mass matrices are also employed. The equations corresponding to each degree of freedom are uncoupled, so only vector operations are necessary.

Explicit finite element codes for commercial use have only recently become available, for example in the forms of MSC/DYTRAN and ABAQUS/Explicit, and no research is known which compares the efficiency of these schemes to the more established implicit codes for solution of low-velocity impact problems. It is established that explicit schemes are preferable for high-speed collisions such as bird strikes on airplanes or ballistic impacts, and implicit methods are more efficient for low-speed and quasi-static events. Low-velocity impacts on laminated composites fall into a gray area between these two extremes. The low impact velocities suggest the use of implicit methods. However, it is necessary to use a refined element mesh in the contact area for impact problems, because the magnitude of the stress gradients with respect to the spatial coordinates is large. A large number of degrees of freedom, and the fact that the explicit packages often include more sophisticated modeling options specifically designed for impact problems, suggest that explicit codes may be preferable.

Contact law refers to the relationship between the indentation of the impacting body into the target and the transmitted force. The distribution of the contact force on the surface of the target is difficult to calculate. It is theoretically infinite for the initial point impact, but then spreads out over a growing finite area. Strain rate effects and plastic or brittle behavior of both the impactor and the target materials are important, as there have been instances observed in which multiple contacts occur during a single impact event [10]. The force-indentation relationship changes during unloading and reloading.

Analysis of contact between two bodies dates back to the work of Hertz in the late 19th century [11]. Today many of these same solutions form the basis of contact laws for composites. The basic form of the results are:

$$F=ka^n$$

F = contact force

k = contact stiffness

a = indentation

n = constant

Unfortunately, composite materials have characteristics which often invalidate the assumptions necessary to solve the equations analytically. A composite lamina is highly orthotropic, and exhibits strain-rate-dependent effects. For most impacts the stresses will be large enough to cause the target material to fail locally like a brittle material so this behavior must be included in the model. Friction between the impactor and target on the surface adds tangential forces.

For these reasons experimental force-indentation data do not agree well with the classic Hertzian law. Researchers have thus turned to experimentally determined contact laws for purposes of modeling. A relationship developed by Tan and Sun [12] has become widely accepted as accurate and is frequently cited in publications by other authors. It is assumed that during the initial loading of an undamaged laminate force versus indentation obeys the Hertzian relationship with a nonlinear (3/2) power. The stiffness of the laminate is unique to each material and stacking sequence and is determined by fitting a curve to experimental data. Upon unloading, the relationship is different. The force is a function of the maximum force during the loading phase, indentation depth and the critical indentation. Critical indentation defines the depth at which permanent deformation begins and must be determined by experiment for each material and stacking sequence.

For certain ratios of impactor mass/velocity and target stiffness, multiple contacts occur. The impactor transfers the majority of its momentum to the laminate, which then breaks contact with the now relatively slow-moving impactor. As the target rebounds from its maximum displacement (it essentially vibrates freely), it re-contacts the impactor before the impactor bounces clear of the target. The contact law has yet a third relationship for this reloading phase. It was found that the 3/2 power relationship no longer holds. Again the exponent must be found experimentally, and often 2 or 5/2 fits the data better. Results from tests such as these are frequently used by other investigators as the contact law of choice in their models. The obvious disadvantage of this method is that the tests require expensive equipment, expertise, and large amounts of time to perform. For every combination of impactor and target material (including stacking sequence) and geometry, a separate test is required. Investigators who do not have the facilities available to perform their own tests are limited to scarce published data.

Somprakit and Huston have developed a numerical method for determining displacements, stresses, and distribution of contact force between contacting cylinders [13]. It is an iterative numerical procedure based upon fundamental solutions from the theory of elasticity. In this technique the cylinders are of infinite length; it is a 2-D analysis. The contact surface is discretized in the x (transverse) direction, and the stresses are found in the x and z (depth) directions. The problem of a spherical impactor on a flat plate can be solved in an analogous manner. The contact area, rather than being defined by its width, is defined by its radius. For now, oblique impacts are not considered. In the Somprakit and Huston's analysis, the infinite length of the cylinders allows the problem to be simplified to one of two dimensions. In the impact problem, axisymmetry is assumed to simplify the equations.

Knowledge of the relationship between the indentation of the impacting body in the target and the magnitude and distribution of the contact force is essential for accurate modeling and determination of the stresses which cause damage. A contact law which predicts total force without the details of its distribution may be sufficient to model the gross deformations of the target under impact loading, but will not allow a model to be as accurate in predicting the stresses which cause the matrix cracking and delaminations representative of low-velocity impact.

Damage prediction is an area which many researchers have made considerable efforts. For a given state of stress in a laminate, a model must be capable of predicting the onset and propagation of damage. Matrix cracking, ply delamination, and fiber breakage are the normal sequence in which damage takes place during a low-velocity impact event. Properties of the laminate will change as damage occurs, so the model must be continually updated (in a temporal sense) as the analysis proceeds. Knowledge of the material's behavior beyond the elastic range and its fracture characteristics are needed. The damage modeling must deal with failed material (cracked matrix and broken fibers) by either removing it or treating it with revised properties, and also with changing laminate properties.

A recent paper by Choi and Chang of Stanford [3] proposes two mechanisms by which delaminations occurs. Both are the result of matrix cracking. In the inner layers a "shear matrix crack" generates delamination which propagate unstably. A small stable delamination occurs at the interface above the cracked layer, and the larger unstable delamination occurs at the interface below. The delamination is governed by the interlaminar shear stress in the fiber direction of the layer below ($\sigma_{xz}$) and the interlaminar shear stress in the direction normal to the fibers ($\sigma_{yz}$) in the layer above the delaminated interface. The critical matrix crack is in this upper layer.

In the layers toward the bottom "bending matrix cracks" occur and stable delamination at the interface is seen. Again the delamination is governed by the transverse shear stress in the fiber direction ($\sigma_{xz}$) in the layer below the interface but it in the upper layer it is the transverse in-plane stress normal to the fibers ($\sigma_{yy}$) which contributes.

For either type of delamination, matrix cracking is the event which triggers delamination. So, failure criteria incorporating the stresses named above are presented to predict matrix cracking and subsequently the positions and sizes of the delaminations. Experimental data is presented to verify the model.

An interesting analytical theory has been developed by Liu [4] which predicts the oft-seen lemniscular or "peanut" shaped delaminations associated with low velocity impact. This is based on the quantity known as the mismatch angle, which is the difference in fiber orientation between adjacent lamina in a laid-up composite structure. A mismatch coefficient is derived from the difference in bending stiffnesses of the upper and lower lamina at an interface which predicts the relative area and orientation of the delamination. For an interface where the adjoining layers are oriented in the same direction, no delamination is predicted, which is consistent with experimental evidence that delamination only exists where the fibers change orientation. The effects of material properties, laminate thickness, and impact energy are discussed.

A great amount of test data has been gathered and published by researchers but there is yet to be an accepted measure of impact performance that is independent of the test method details [5]. The variations in impactor mass and size, impactor velocity, test specimen size and boundary, et cetera are so wide that no coordinating theory has been derived to predict the behavior for any given set of impact parameters.

The purpose of damage prediction is to be able to design composite structures which are tolerant to foreseeable impact events. Low velocity impacts an aircraft might experience include tool drop, hail, footsteps, and runway debris. If damage can be predicted, then the properties of the composite structure post-impact can be characterized also in order to estimate the residual strength and stiffness. Laminates may suffer a significant degradation of their properties due to low velocity impact, and to design structures that can survive such impacts, one must be able to predict the damage which is likely to occur.

The determination of residual properties has been undertaken by many investigators. The strength of damaged laminates, particularly in compression, is an important field of study. Since low velocity impact damage can extend such a relatively long distance from the impact site, strength can be reduced much more than for the case of a penetrating impact. Similarly stiffness can be change significantly due to back-ply damage.

Frequently there is a loss of symmetry in a laminate due to impact damage. This introduces bending-stretching coupling. The vast majority of composite laminates are laid up symmetrically about the mid-plane in order to eliminate coupling of the bending and extensional strains. This greatly simplifies the analysis of such laminates. But when damage occurs the symmetry is lost and the behavior of the laminate may change drastically.

The effects of impact damage on tensile strength has been investigated by El-Zein and Reifsnider [14]. They hypothesized that residual strength is controlled by stress concentration effects in the immediate vicinity of the damaged region. A complex variable solution based on Lekhnitskii's problem of an anisotropic plate with an elliptical inclusion is derived, and the results agree fairly well with experiment.

The change in compression strength after impact is a more widely reported phenomenon. Dost, Ilcewicz, and Gosse presented a sublaminate stability based approach to predict damage tolerance (i.e. post-impact strength) [15]. This approach does require an accurate description of the state of the damage inside the plate. Further it was noted that there were significant differences in residual strength between experimental coupons and actual composite structures due to finite width effects in the test specimens.

Buckling of delaminated composites is a third failure mode. Global buckling is the same mode as occurs in undamaged panels, but may occur at markedly smaller load levels when damage is present due to loss of stiffness and the previously mentioned bending-stretching coupling. The phenomenon of local or delamination buckling is strictly related to delaminations near the free surface of the laminate. Under load the delaminated region may buckle while the rest of the laminate remains stable. Chai and Babcock [16] modeled an anisotropic layer separated from a thick isotropic base laminate. The delamination is elliptic in shape and the material axes coincide with the ellipse's axes. The buckling for the damaged region by the Rayleigh-Ritz method and propagation of the delamination area is predicted via a fracture mechanics approach. Results show stable, unstable, or unstable growth with crack arrest depending on material properties and orientation, loading, and fracture energy.

Davidson [17] considered failure of a damaged laminate by all three failure modes: compression, global buckling, and delamination buckling. Buckling loads are calculated by applying the Trefftz criterion to governing equations found from the Rayleigh-Ritz method and compression failure by a modified maximum strain criterion. If the initial failure is delamination buckling, that layer is removed (it carries no load), the laminate properties are recalculated, and the loading sequence is continued until catastrophic failure (compressive or global buckling) is reached.

Davidson compared five analyses (two performed by himself and three reported by other researchers including Chai and Babcock) to experimental results. He found only one gave conservative predictions of buckling loads. The remaining four over-estimated the failure loads. The model which gave conservative results employed the reduced bending stiffness approximation. The [D] matrix is replaced by a matrix $[D^*]$ defined as $[D^*]=[D]-[B][A]^{-1}[B]$ for cases where the coupling matrix [B] is non-zero. The analysis is then performed as though the laminate was symmetric.

It was discovered that under certain conditions delamination buckling can occur under tensile loading. The fibers in a lamina normally have a small Poisson's ratio, approximately one order of magnitude smaller than the matrix material or a quasi-isotropic laminate. When the fibers in a delaminated layer are oriented normal to a tensile load, both the delamination and the base undergo lateral contraction. Due to the mismatching of the Poisson's ratios, the delaminated region experiences compression and the base tension. Thus buckling may occur under loading cases when it is not expected.

Of these five areas that were reviewed, those concerning the dynamic response of laminated plates and residual strength of damaged composites seem to be the furthest developed. Plate theory was a topic of great interest even before the development of composites. The extension of research on these materials has paralleled their increasing use in engineered structures. Residual strength models have not received the same volume of attention, but researchers have been able to demonstrate the capability to predict post-impact characteristics for composite laminates with known, although admittedly simplistic, damage states. Damage prediction has recently been the focus of several researchers. Unfortunately the state of the art is not as advanced as for the first two topics. Several different theories have been proposed to predict the onset of damage at a specified level of stress in a laminate. The focus of this effort has been in the two remaining categories, contact law and numerical modeling of impact events.

Contact law appears to be an area which has not received as much attention from researchers as the other aspects of impact analysis, but there are still questions which remain. New numerical techniques, particularly from the finite element community, need to be investigated to determine their applicability to impact analysis. A new method for analytically developing a contact law has been the major thrust of the work. The advantage this new technique will bring is that it will predict the distribution of contact pressure between an impactor and a target, not just the total force. Current relationships only calculate the total force versus indentation. When incorporated into models in which the contact region is discretized into numerous sub-regions, such methods cannot be used without subsequent assumptions concerning the allocation of the force.

The advancement of numerical analysis techniques which will take full advantage of such a contact law naturally follows. As computing speeds and capabilities grow in a general sense, development of specific methods for solving impact problems deserves attention.

The contact law under development is based upon the theory of elasticity. For cases involving torsionless axisymmetry, solutions can be found for a host of loading and boundary conditions [18]. By a series of derivations the problem of a constant distributed load over a finite radius acting on an isotropic half space can be solved. This pressure-displacement relationship is the basis for the contact law. The solutions show that stresses diminish rapidly at distances a few contact radii from the original contact point, so the assumption of a semi-infinite target to represent a plate will give more accurate results for thicker laminates. Axisymmetry is utilized to reduce the problem from that of three dimensions to two. Thus, the contact surface can be modeled as a single radial line. Since the surface ply of a composite laminate is orthotropic, modifications must be made to account for the assumption of isotropy. The equations upon which the analysis is based follow:

$$\delta = \frac{4(1-v^2)P\rho}{\pi E}\int_0^{\pi/2}\sqrt{1-\left(\frac{r}{\rho}\right)^2\sin^2\Phi d\Phi} \quad \text{, for } r \leq \rho$$

$$\delta = \frac{4(1-v^2)Pr}{\pi E}\left[\int_0^{\pi/2}\sqrt{1-\left(\frac{\rho}{r}\right)^2\sin^2\Phi d\Phi} - \left\{1-\left(\frac{\rho}{r}\right)^2\right\}\int_0^{\pi/2}\frac{d\Theta}{\sqrt{1-\left(\frac{\rho}{r}\right)^2\sin^2\Theta}}\right] \quad \text{, for } r > \rho$$

P = constant pressure

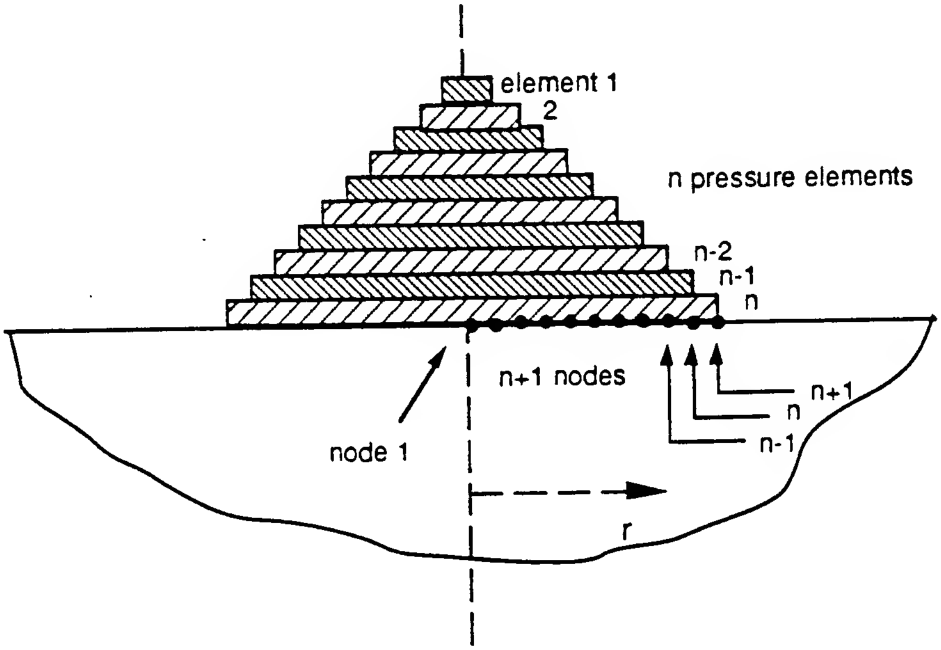
ρ = radius of area over which pressure acts

E,v = material properties of half-space

δ = transverse deflection

r = radial coordinate

The contact area is represented by N overlapping constant pressure elements extending radially from the axis of symmetry. There is one node common to all elements at the center of the area. The second node of each element discretizes the contact area into regularly spaced sub-radii. Each node has a single translational degree of freedom in the transverse direction. Influence coefficients relating the elemental pressures to the nodal displacements are derived from the elasticity equations. A set of initial displacements is determined based on the user-specified parameters of the problem.



At the end of each iteration, the validity of the solution is checked based on three criteria:

1) Equilibrium

   The sum of the elemental pressures multiplied by their corresponding areas must equal the applied force.

2) Edge Pressure

   The pressure in the outermost element must approach zero at the edge of the contact zone. Since the contact area is divided into discrete areas, the pressure in this element cannot equal zero, or else there would be no contact.

3) Contact/Separation

There must be continuous contact between the target and impactor in the assumed contact region and separation outside the contact region. For the geometry chosen, a spherical impactor on a flat target, contact cannot be broken at a given radius and re-established at a greater radius.

When these three conditions are satisfied a solution has been found and iteration stops. It was found that satisfying all three conditions simultaneously is difficult due to the non-linearity of the problem. Specifically, the size of the contact area varies rapidly with changing load.

The original solution method relied on a heuristic approach to solving the equations. But the rate of convergence depended very strongly on the initial estimate of the displacements. In an effort to find more accurate solutions, optimization methods were investigated.

Optimization refers to any of a number of algorithms which seek to minimize a mathematical function, called the objective function. In addition, extra requirements to be satisfied, known as constraints, may be placed on the solution. The parameters from which the objective function and constraints are constructed are the design variables. Although mathematical in theory, optimization techniques have been of great interest to engineers as a means of solving difficult systems of non-linear equations which frequently arise, or for finding the "best" solution to a problem of many variables which has no exact solution. If the important physical phenomena in an engineering system can be accurately described by an objective function and constraints, a numerical method can often be found which will provide satisfactory solutions when analytical methods do not.

For the contact problem, several equations have been derived relating surface pressure and displacement. It is desirable to use these equations as the basis for the objective function and constraints, since they are known to govern the physical system. For a similar contact problem, optimization techniques were applied to the Rayleigh-Ritz method [19]. Rayleigh-Ritz is an energy method which requires a mathematical function to describe the internal strain energy of the elastic body. It was noted that it is difficult to find a function which satisfies the kinematic boundary conditions for a contact problem and gives reasonable solution accuracy.

The first approach is to make the objective function by rearranging the equilibrium equation. The optimization routine would seek to minimize the absolute value of the difference between the sum of the pressures multiplied by their respective areas and the applied force. The conditions of the edge pressure approaching zero and the contact/separation condition would be imposed as constraints. A second approach would be to use the edge pressure criterion as an objective function, and the other two as constraints.

The computer modeling research focused on using a finite element employing an explicit time integration scheme for solution of the equations of motion. Specifically, the DYNA3D code developed at Lawrence Livermore National Laboratories was used. The Ohio Supercomputer Center made available its Cray Y-MP 8/264 computer for this phase of the research. Dr. David Lemmon of the University of Cincinnati was instrumental in this effort, both in obtaining the grant of the resources and his technical expertise.

DYNA3D has considerable capabilities for solving engineering problems, but it lacks a pre-processor for building finite element models. The I-DEAS package developed by Structural Dynamics Research Corporation was used for pre-processing. Dr. Lemmon provided a translator which converts data stored in an I-DEAS universal file to a DYNA3D input deck.The case that was chosen to be modeled was a spherical impactor dropped on a composite plate. This event has been analyzed by many researchers and can be most easily recreated in a laboratory experiment. Although the geometry of the model is uncomplicated, it incorporates several features which are necessary in order for DYNA3D to be able to solve the problem, such as compatibility of element size on the contact surfaces of the impactor and target. Without this precaution, a valid input deck will be written but fatal errors will be encountered during the solution, wasting computer time and necessitating correction of the model. It is also necessary to specify entities in the I-DEAS model which will be used to define sliding interfaces, initial velocities, and material models in

DYNA3D. The translator writes the vast majority of the input cards for the analysis. It transfers all the geometry, nodes, and elements directly and allows the user to easily create the sliding interfaces, initial conditions, and material models if the appropriate groups of nodes and elements are contained in the universal file.

The DYNA3D input deck must be edited to alter the job control cards at the beginning of the deck and to edit the material cards to include the composite. The translator does not handle composite materials. Individual ply properties and stacking sequence must be input, and a user-defined integration rule defining the number of through-the-thickness integration points must be specified. Damage is included in the material model. For solid elements, one of a number of widely-recognized composite failure theories can be chosen. For shell elements, which were used for this analysis, individual failure theories detecting matrix cracking, matrix crushing, ply delamination, and fiber breakage are checked at each time step. If failure by any mode is predicted, the material properties associated with that failure in the particular elements are reduced to zero over the next one hundred time steps in order to reduce numerical difficulties stemming from the sudden release of strain energy. At this stage of the research however, damage was not to be considered so the strengths of the material were made artificially high to prevent failure.

The work using DYNA3D was undertaken only recently and is ongoing. At present, there has been success in the modeling phase and in performing analyses for isotropic materials, such as a steel impactor striking a steel plate. The companion post-processor to DYNA3D, called TAURUS, can be used to display results, and it is also possible to import the results into a more sophisticated post-processing program such as I-DEAS. But when a composite plate is analyzed, the solution encounters numerical difficulties during the analysis and aborts. It is believed that this is due to problems with the sliding interfaces.

Sliding interfaces are specified in DYNA3d to define contact surfaces on bodies which cannot penetrate one another. There are several assorted types of interfaces available which use different mathematical formulations to enforce the contact conditions. In general groups of nodes and/or elements are chosen to define master and slave surfaces which describe the contact surfaces.

Two type of interfaces are considered for this problem, both penalty formulations. The "sliding with separation and friction" option is most general. It includes friction and allows the two specified surfaces (shell elements of the target plate, surface faces of the solid elements of the impactor) to contact and separate arbitrarily. It does not matter which of the two surfaces is selected as the master and slave. For the "discrete nodes impacting surface" option, the slave group is not a surface, but rather a set of nodes [20].

Each of the available types of sliding interfaces has its own idiosyncrasies. Some are more robust than others so some experimentation is necessary to determine which are compatible with the composite materials.

# CONCLUSIONS

During the Summer Research Extension Program a thorough review of existing research in several areas concerning low-velocity impact of composite materials was performed. Based on this review, the focus of the new research was chosen to be the topics of contact law and numerical modeling. A new technique for determining the contact force-indentation relationship for low-velocity impact of composite materials has been under development. This contact law is critical for modeling of impact events. Present methods require experimental results to determine several parameters. Different combinations of impactor and target materials require separate tests. Due to the non-linearity of the contact phenomenon, it is very difficult to extrapolate results or predict behavior when no experimental results are available.

The new method also gives information about the distribution of the contact pressure underneath the impactor. This is particularly useful for performing finite element analyses where the contact forces must be applied at several discrete points. The stress gradients are extremely high in the contact zone, so in order to accurately model impact and be able to track the stresses which initiate damage, this type of precision is necessary. The contact laws which are currently used do not provide such information. A computer program is under development which seeks to reduce the amount of experimental data necessary for modeling of impact events. The method is based upon equations derived from the theory of elasticity. An experimental plan for validating the theoretical aspects of this method will be proposed.

Numerical methods were investigated to study impact and contact problems. These include modeling for commercially available codes such as ANSYS, which solves transient dynamics problems with a conventional implicit solution scheme, and DYNA3D, which uses an explicit scheme to solve the system of equations. The explicit solution method is advantageous for problems involving high-velocities and/or large deformation. For a low-velocity impact problem, it is not clear which method will have the advantage in computation cost.

## REFERENCES

1. Finn, S.R. and Springer, G.S., Delaminations in composite plates under transverse static loads - A model. Composite Structures, 23 (1993) 177-190

2. Finn, S.R. and Springer, G.S., Delaminations in composite plates under transverse static loads - Experimental results. Composite Structures, 23 (1993) 191-204

3. Choi, H.Y. and Chang, F.K., A Model for Predicting Damage in Graphite/Epoxy Laminated Composites Resulting from Low-Velocity Point Impact. Journal of Composite Materials, 26 (1992) 2134-2169

4. Liu, D., Impact-Induced Delamination - A View of Bending Stiffness Mismatching. Journal of Composite Materials, 22 (1988) 674-692

5. Robinson, P. and Davies, G.A.O., Impactor Mass and Specimen Geometry Effects In Low-Velocity Impact of Laminated Composites. International Journal of Impact Engineering, 12 (1992) 189-207

6. Noor, A.K., and Burton W.S., An Assessment of Computational Methods For Multi-Layered Shells. Applied Mechanics Reviews, 43 (1990) 67-97

7. Lee, J.D., Du, S., and Liebowitz, H., Three-Dimensional Finite Element and Dynamic Analysis of Composite Laminate Subjected to Impact. Computers and Structures, 19 (1984) 807-813

8. Sun, C.T. and Chen, J.K., On the Impact of Initially Stressed Composite Laminates. Journal of Composite Materials, 22 (1988) 490-504

9. Wu, H.T. and Springer, G.S., Impact Induced Stresses, Strains, and Delaminations in Composite Plates. Journal of Composite Materials, 22 (1988) 533-560

10. Sun, C.T. and Chen, J.K., "On the Impact of Initially Stressed Composite Laminates", Journal of Composite Materials, Vol. 19, November 1985, pp. 490-504

11. Hertz, H., Uber die Behrurung fester Elastischer Korper. Journal Reine Angle Math, 92 (1881) 155

12. Tan, T.M. and Sun, C.T., Use of Statical Indentation Laws in the Impact Analysis of Laminated Composite Plates. Journal 6 Applied Mechanics, 52 (1985) 6-12

13. Somprakit, P., Huston, R.L., and Wade, J.E.II, Monitoring of Contact Stresses In Advanced Propulsion Systems. University of Cincinnati, Cincinnati, Ohio, 1990

14. El-Zein, M.S., and Reifsnider, K.L., "On the Prediction of Tensile Strength after Impact of Composite Laminates", Journal of Composites Technology and Research, Vol. 12, No.3, 1990, pp. 147-154

15. Dost, E.F., Ilcewicz, L.B., and Gosse, J.H., "Sublaminate Stability Based Modeling of Impact-Damaged Composite Laminates", Proceedings of the American Society for Composites, 3rd Technical Conference, Seattle, WA, 1988, pp. 354-363

16. Chai, H., and Babcock, C.D., "Two-Dimensional Modeling of Compressive Failure in Delaminated Laminates", Journal of Composite Materials, Vol. 19, January 1985, pp. 67-98

17. Davidson, B.D, "A Determination of the Strength and Mode of Failure of Compression Loaded Laminates Containing Multiple Delaminations", JPL Document D6447, September 1989

18. Timoshenko, S.P. and Goodier, J.N., Theory of Elasticity, 3rd Edition, McGraw-Hill, 1970

19. McDonald, E.S., Optimization Techniques For Contact Stress Analysis, M.S. Thesis, Naval Postgraduate School, Monterey, California, 1992

20. Whirley, R.G., DYNA3D User's Manual, University of California, Lawrence Livermore National Laboratory, 1991

# DETECTION OF INTERNAL DEFECTS IN MULTILAYERED PLATES BY LAMB WAVE ACOUSTIC MICROSCOPY

Tribikram Kundu
Associate Professor
Department of Civil Engineering
and Engineering Mechanics

University of Arizona
Tucson, Arizona 85721

Final Report for:
Summer Research Extension Program
Wright-Patterson Material Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

University of Arizona

December 1993

# DETECTION OF INTERNAL DEFECTS IN MULTILAYERED PLATES BY LAMB WAVE ACOUSTIC MICROSCOPY

Tribikram Kundu
Associate Professor
Department of Civil Engineering and Engineering Mechanics
University of Arizona

## Abstract

Under this research contract a theoretical study has been carried out that shows an excellent potential of detecting small internal defects in multilayered plates by Lamb Wave Acoustic Microscopy (LAM) technique. Unlike conventional scanning acoustic microscope (SAM) which are commercially available, the Lamb wave acoustic microscope is still a concept. SAM generates Rayleigh waves to produce the acoustic image of the near surface defects in a specimen. The Rayleigh wave has a poor penetration property specially at high frequencies since it penetrates only about one wave length deep into a material. Hence, at high frequencies, although its resolution is high, SAM cannot detect relatively deeper cracks in a material. Lamb waves (also known as the "plate waves") on the other hand propagate through the entire plate. For a multilayered plate various Lamb wave modes excite different levels of energy in different layers. A number of Lamb wave modes in multilayered plates have been studied from which it is concluded that different Lamb wave modes can be used to detect defects in different layers.

# DETECTION OF INTERNAL DEFECTS IN MULTILAYERED PLATES BY LAMB WAVE ACOUSTIC MICROSCOPY

Tribikram Kundu

## INTRODUCTION

Increasing use of multilayered materials in engineering applications warrants accurate detection of small cracks inside a layer for safe operations. Ultrasound is often used for this purpose because of its ability to penetrate into the material. However, detecting small material defects inside a multilayered or multiply composite plate is often not an easy task. In such multilayered specimens surface and near-surface-defects can be easily detected by conventional optical techniques or scanning acoustic microscopes (SAM). In SAM, interference between near surface defects and surface skimming Rayleigh waves produces the image of the defects. Thus very small near surface defects can be detected by SAM. The only shortcoming of SAM is that it cannot detect deeper defects, because Rayleigh waves do not penetrate deep into the material.

Deeper defects are generally detected by conventional ultrasonic techniques analyzing the signals which are reflected by the defects (cracks and delaminations). For multilayered specimens this is not an easy task because reflected signals also arrive from material interfaces. Even if these signals can be identified and separated, real difficulties arise when defects in the deeper layers lie in the shadow of other defects as shown in Figure 1, crack 2. Ultrasonic technique also fails to detect defects which are located parallel to the wave propagation direction, crack 4 in Figure 1, since they do not reflect the signal back to the transducer. Because cracks 2 and 4 are located deep inside the material SAM also fails to detect them. These types of difficult-to-detect-defects can be

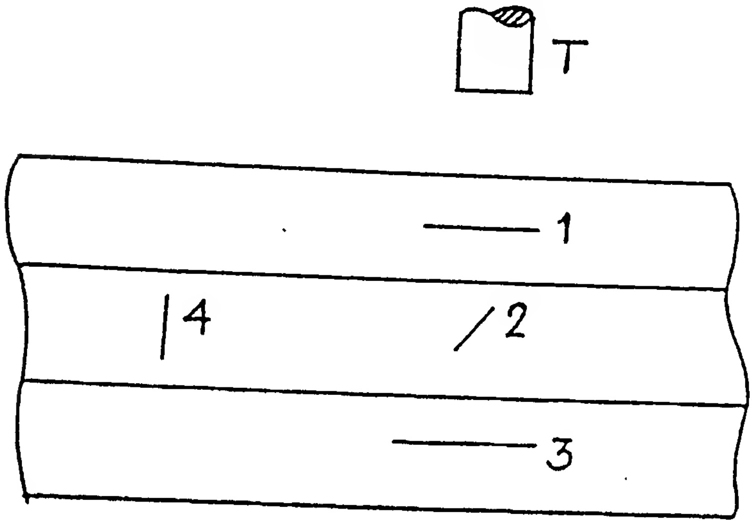


Figure 1: A multilayered plate specimen with internal cracks of different orientations.

detected by Lamb waves as discussed in this report.

For this purpose, first, different Lamb wave modes are to be generated in a layered plate. From the mechanics of elastic wave propagation in multilayered plates one can get the frequencies and angles of incidence for producing different Lamb wave modes. These modes are then studied carefully to see which mode produces strong excitation in which layer.

## PROBLEM FORMULATION

As mentioned above one needs to study the mechanics of elastic wave propagation in multilayered elastic plates to decide which Lamb wave mode is appropriate for generating strong excitation in a specific layer. The fundamental problem of elastic wave propagation in a multilayered solid has been studied by a number of investigators [1-11]. Here, a brief description of the transfer matrix formulation is given and the modifications necessary for avoiding numerical difficulties are pointed out.

The geometry of the multilayered plate is shown in Figure 2. The plate contains N homogeneous layers with perfect bonding at the interfaces. The thickness, P and S wave speeds, and the density of a general m-th layer which lies between the interfaces $z_{m-1}$ and $z_m$ are denoted by $h_m$, $\alpha_m$, $\beta_m$ and $\rho_m$, respectively. The corresponding properties of the fluid are denoted by the subscript f. A time harmonic plane acoustic wave of unit amplitude and circular frequency $\omega$ is incident on the fluid-solid interface at an angle $\theta$. We are interested in computing the reflected field (R), the transmitted field (T) and the wave amplitudes $a_m$, $b_m$, $c_m$ and $d_m$ in the m-th layer.

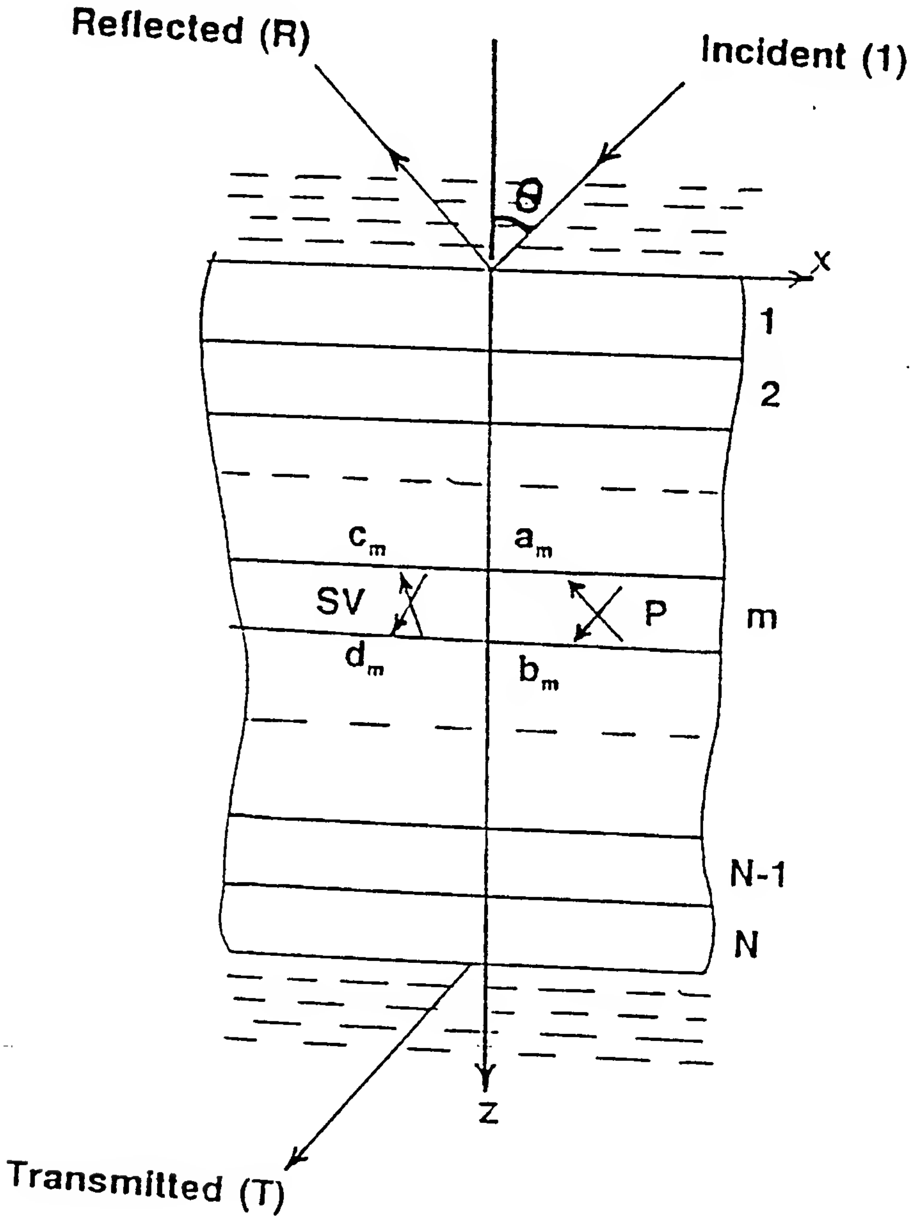


Figure 2: A multilayered plate in a fluid with incident, reflected and transmitted waves

The wave potentials in the m-th layer are given by $\phi_m(z)\exp(-ikx-i\omega t)$ and $\psi_m(z)\exp(-ikx-i\omega t)$, where

$$\begin{aligned}\phi_m &= a_m e^{-iq_m(z-z_{m-1})} + b_m e^{iq_m(z-z_{m-1})} \\ \psi_m &= c_m e^{-ig_m(z-z_{m-1})} + d_m e^{ig_m(z-z_{m-1})}\end{aligned} \tag{1}$$

and the wave potentials in the fluid are $\phi_o(z)\exp(-ikx-i\omega t)$ at the top fluid half space and $\phi_T(z)\exp(-ikx-i\omega t)$ at the bottom fluid half space, where

$$\begin{aligned}\phi_o &= e^{iq_f z} + Re^{-iq_f z} \\ \phi_T &= Te^{iq_f(z-z_N)}\end{aligned} \tag{2}$$

In equations (1) and (2), $a_m$, $b_m$, $c_m$, $d_m$, R and T are unknown constants to be determined from the interface and boundary conditions. These conditions can be satisfied using Thomson-Haskell matrix method [1,2]. By this technique the stress-displacement vector at the top interface is related to the stress-displacement vector at the bottom interface in the following manner

$$S_n = JS_o \tag{3}$$

where

$$J = A_N A_{N-1} \cdots A_1 \tag{4}$$

and $A_m$ is the 4x4 layer matrix, also known as the propagator matrix, transfer matrix or T matrix [1,2,4,8,10,11].

The stress-displacement vectors $S_o$ and $S_N$ at the top and bottom surfaces of the plate are given by

$$S_o=[u_o \quad iq_f(1-R) \quad -\rho_f\omega^2(1+R) \quad 0]^T$$
$$S_N=[u_N \quad iq_fT \quad -\rho_f\omega^2T \quad 0]^T \qquad (5)$$

Inserting these expressions in equation (3) reduces it to a system of four algebraic equations for the four unknowns R, T, $u_o$ and $u_N$. After obtaining R and T the wave amplitudes $a_m$, $b_m$, $c_m$ and $d_m$ at any general m-th layer can be obtained by carrying out the Thomson-Haskell matrix product between the m-th layer and top or bottom surface of the plate [8].

Numerical Difficulties

Thomson-Haskell matrix method has inherent numerical problem of loss of precision. The problem was first noticed by Dunkin [3]. He indicated that one can avoid this numerical difficulty to a great extent by working with the subdeterminants of the transfer matrix. This procedure is called the delta-matrix method. Since then, other schemes have been described, all of them leading to formulations that do not really ensure numerical stability [5-7]. Later Kundu and Mal [8] identified a second precision problem that occurs when computing the amplitude of the transmitted signal and suggested some alternatives to the conventional delta matrix technique to improve the precision problem. Finally, Levesque and Piche [10] made more improvements to the delta matrix method to get rid of all precision problems associated with the elastic wave propagation analysis in multilayered solids.

## NUMERICAL RESULTS

To illustrate the feasibility of using different Lamb wave modes to detect defects in different layers in a multilayered solid two different types of specimens with four different geometries are considered. The first type specimen is a two layered plate made of copper and aluminum. The second type is a three layered plate specimen made of two aluminum plates glued together by a thin epoxy layer. These specimens are shown in Figure 3. Each type of specimen has two different geometries. For the two layered plate the thickness of the layers is taken as 2 mm (specimen 1a) and 3 mm (specimen 1b). Aluminum plates in the three layered specimen have thickness 3.96 mm and 2.54 mm but the epoxy layer thickness is varied. Specimen 2a has epoxy thickness 0.5 mm and 2b has this thickness equal to 0.1 mm. Table 1 shows the material properties used in the following analyses.

The Lamb wave dispersion curves for the four different specimens (1a, 1b, 2a and 2b) are first computed and shown in Figure 4. From these dispersion curves one can obtain the Lamb wave phase velocities at any frequency. For example, at 1 MHz frequency the four specimens have 5, 6, 7 and 7 different phase velocities between 0 and 10 km/sec. These are shown in Table 2. Three more phase velocities, which are greater than 10 km/sec can be obtained by extrapolating the dispersion curves for three specimens 1b, 2a and 2b. These are shown in parentheses in table 2. For every phase velocity the corresponding critical angles can be computed from Snell's law. These angles are shown next to the phase velocities in Table 2.

Reflection and transmission coefficients are then computed for these specimens as a function of the incident angle θ. These plots are shown in Figure 5, for 1 MHz signal frequency. The continuous and broken curves in each plot represent reflection and

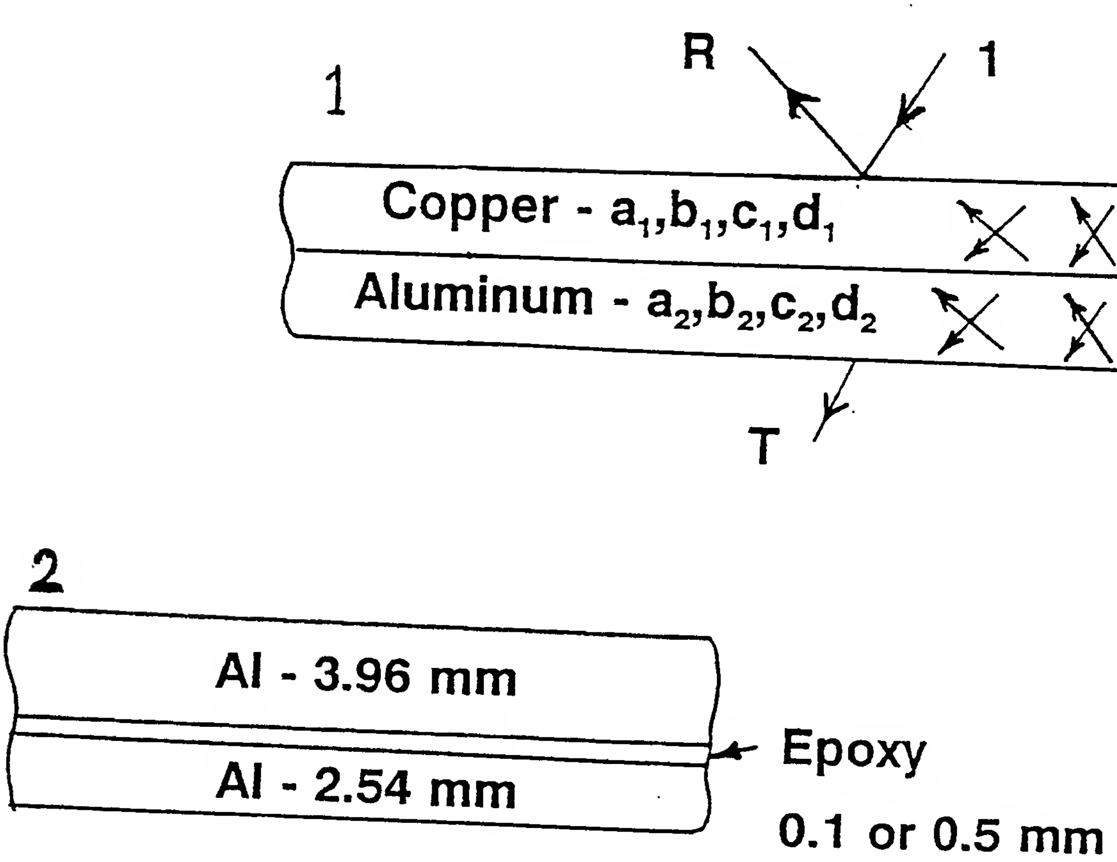


Figure 3: Plate specimens for which results are presented.

Table 1: Table of Material Properties.

| Material | Density | P-Wave Speed | S-Wave Speed |
|---|---|---|---|
| Copper | 8.93 gm/cc | 4.66 km/sec | 2.66 km/sec |
| Aluminum | 2.70 gm/cc | 6.32 km/sec | 3.13 km/sec |
| Epoxy | 1.20 gm/cc | 2.20 km/sec | 1.10 km/sec |
| Water | 1.00 gm/cc | 1.49 km/sec | 0 |

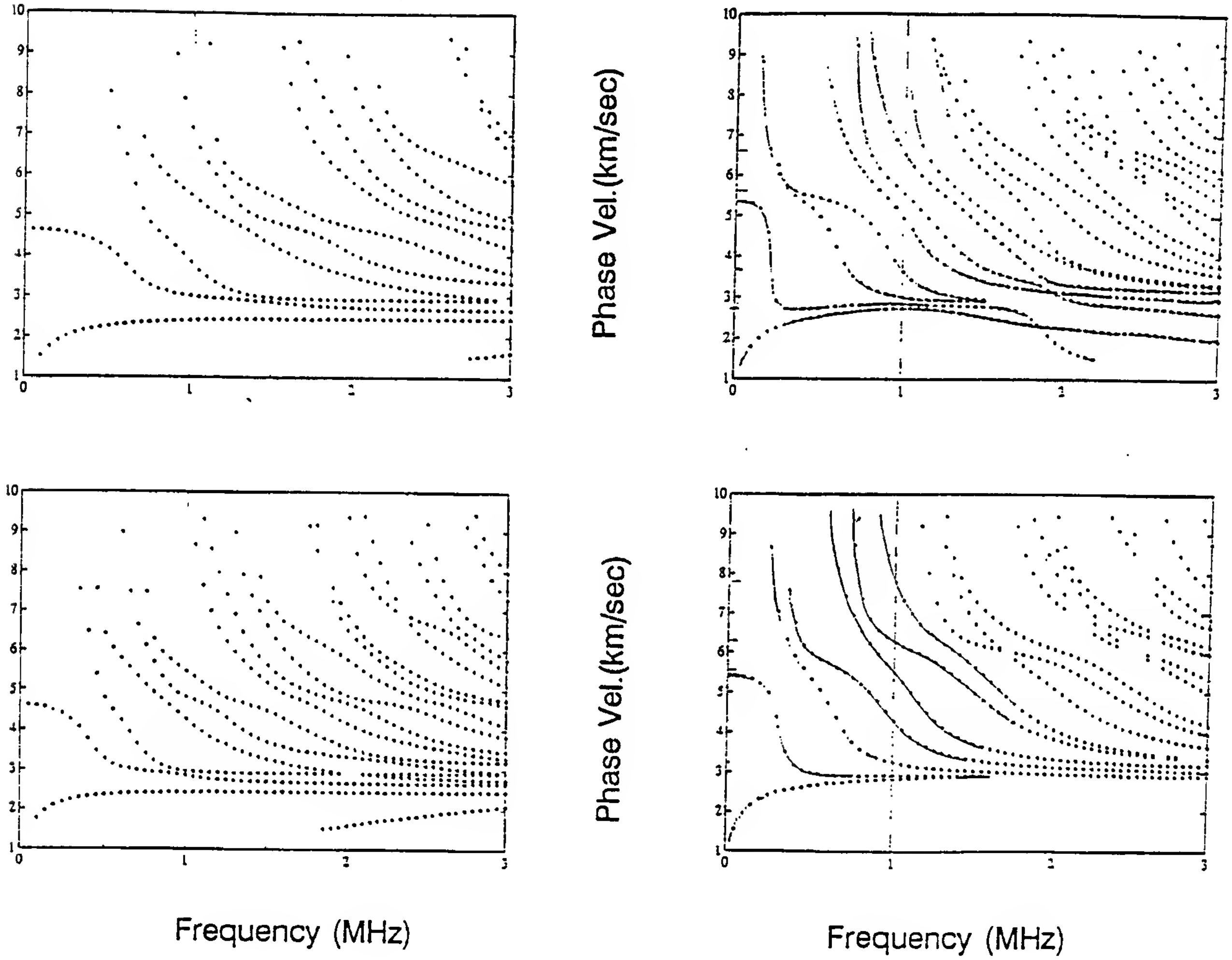


Figure 4: Dispersion curves for specimens 1a (top left), 1b (bottom left), 2a (top right) and 2b (bottom right).

Table 2: Lamb wave phase velocities and critical angles of four specimens at 1 MHz.

2 mm Cu - 2 mm Al

| Ph.Vel. km/sec | Cr.Angl. degree |
|---|---|
| 2.4 | 38.4° |
| 3.0 | 29.8° |
| 3.8 | 23.1° * |
| 5.5 | 15.7° |
| 7.2 | 11.9° * |

1a

3 mm Cu - 3 mm Al

| Ph.Vel. km/sec | Cr.Angl. degree |
|---|---|
| 2.45 | 37.5° |
| 2.9 | 30.9° * |
| 3.0 | 29.8° |
| 4.35 | 20.0° * |
| 4.85 | 17.9° |
| 5.6 | 15.4° * |
| (12.3) | 7.0° * |

1b

Epoxy Bonded Aluminum Plate:

0.5 mm Thick Epoxy

| Ph.Vel. km/sec | Cr.Angl. degree |
|---|---|
| 2.7 | 33.5° |
| 2.8 | 32.1° |
| 3.0 | 29.8° |
| 3.7 | 23.7° |
| 5.0 | 17.3° * |
| 5.6 | 15.4° * |
| 6.65 | 12.9° |
| (12.3) | 7.0° * |

2a

0.1 mm Thick Epoxy

| Ph.Vel. km/sec | Cr.Angl. degree |
|---|---|
| 2.85 | 31.5° |
| 3.95 | 30.3° |
| 3.3 | 26.8° * |
| 4.3 | 20.3° * |
| 5.55 | 15.6° * |
| 6.3 | 13.7° |
| 7.8 | 11.0° |
| (17.1) | 5.0° * |

2b

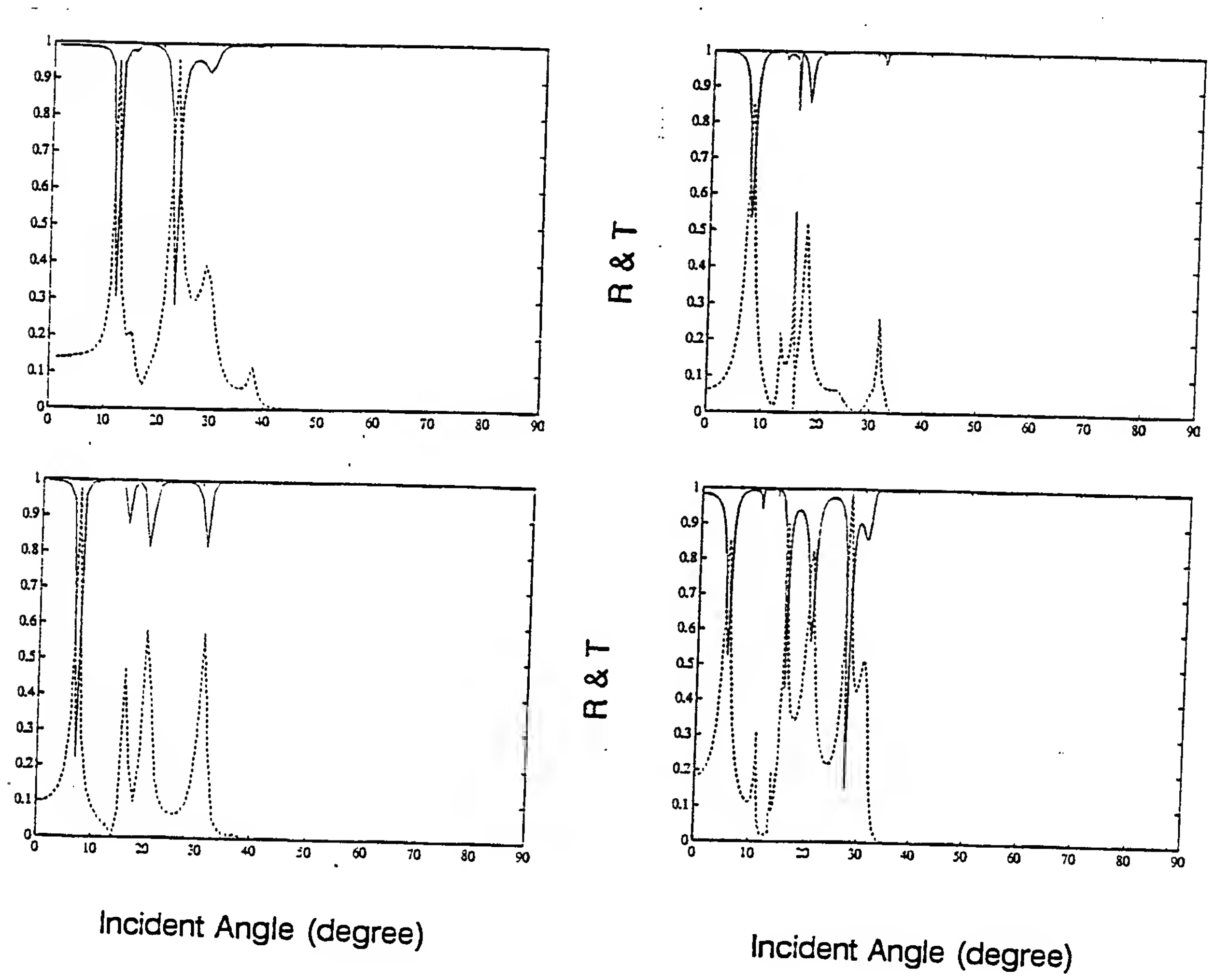


Figure 5. Reflection (continuous) and transmission (broken curves) coefficients of four specimens [1a (top left), 1b (bottom left), 2a (top right) and 2b (bottom right) as the angle of incidence (plotted along the x-axis) varies from $0^{o}$ to $90^{o}$.

transmission coefficients respectively. It should be noted here that the number of dips in the reflection coefficient plots (or peaks in the transmission coefficient plots) are equal or less than the number of phase velocities (or critical angles) at that frequency. Theoretically the reflection coefficient plot should have a dip at every critical angle. If one carefully computes and plots the reflection coefficients against the angle of incidence then one should observe it. One can see from Figure 5 that for specimen 1a there are five critical angles at 1 MHz and there are five dips in the reflection coefficient plot. However, only two of those five dips are strong. The remaining three are weak. Very weak dips can be too weak to observe and be missed. In table 2, the critical angles which show strong dips in the reflection coefficient plots are marked by "*". Strong dips in the reflection coefficient probably indicate relatively strong leaky Lamb wave modes at those critical angles. All four specimens show some strong and some weak dips in their reflection coefficient plots.

Finally the wave amplitudes $a_i$, $b_i$, $c_i$ and $d_i$ ($i = 1$ and 2 for top and bottom layers respectively, see Figure 3) are computed as a function of the incident angle for all four specimens. The results are plotted in Figures 6 and 7. Solid, dashed, dotted and dashed-dotted curves are used to show the variations of $a_i$, $b_i$, $c_i$ and $d_i$ respectively. For specimens 1a and 1b the wave amplitudes at the bottom layer are relatively stronger for $12^o$ and $7^o$ incidence respectively. At larger angles $40^o$ or greater the wave amplitudes at the top layer are relatively stronger. Hence, one can logically conclude that for detecting cracks and defects at the bottom layer in specimens 1a and 1b the transducers (transmitter and receiver) in a pitch-catch arrangement should be positioned at $12^o$ and

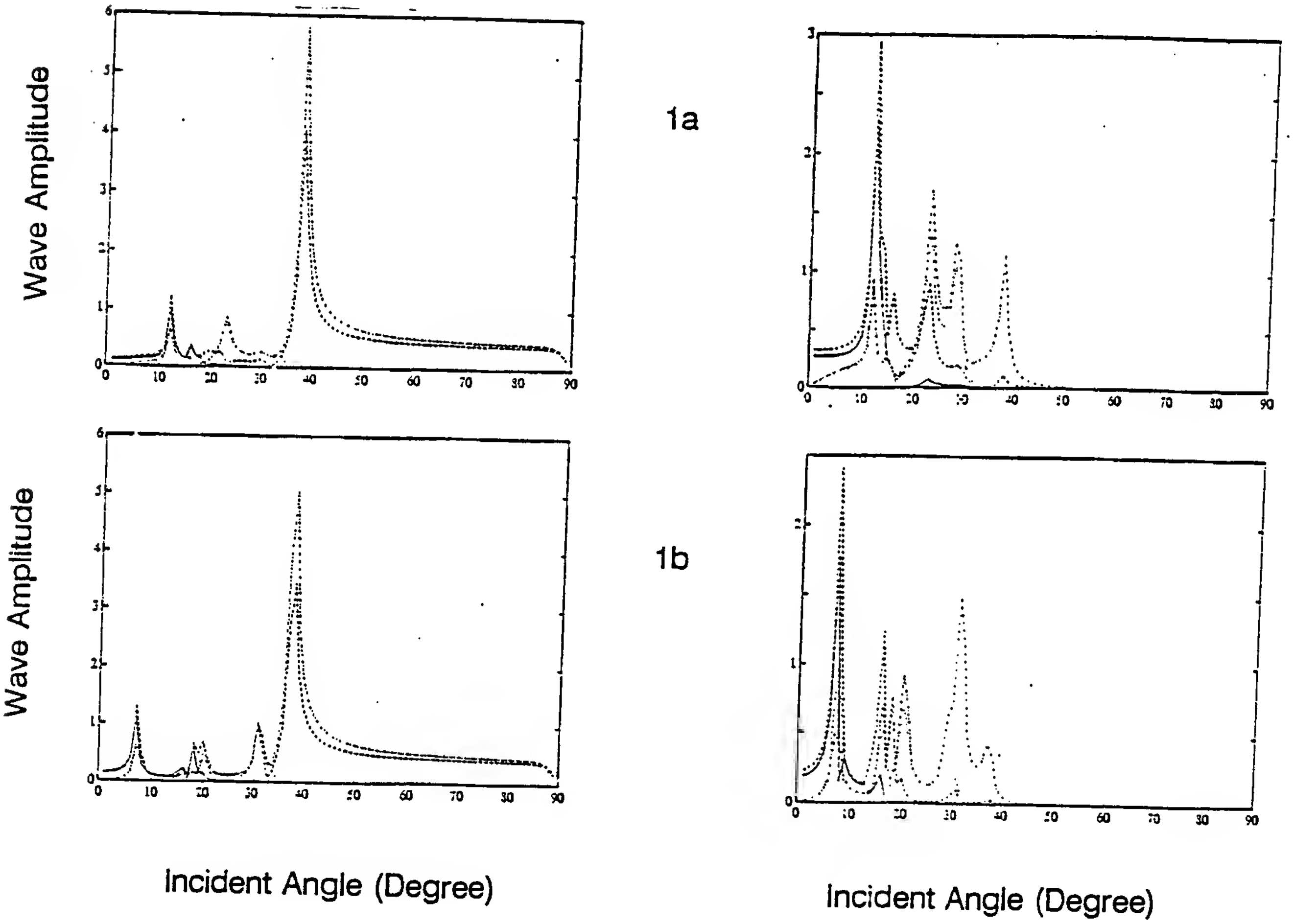


Figure 6: Wave amplitudes in top layer (left column) and bottom layer (right column) of specimens 1a (top row) and 1b (bottom row).

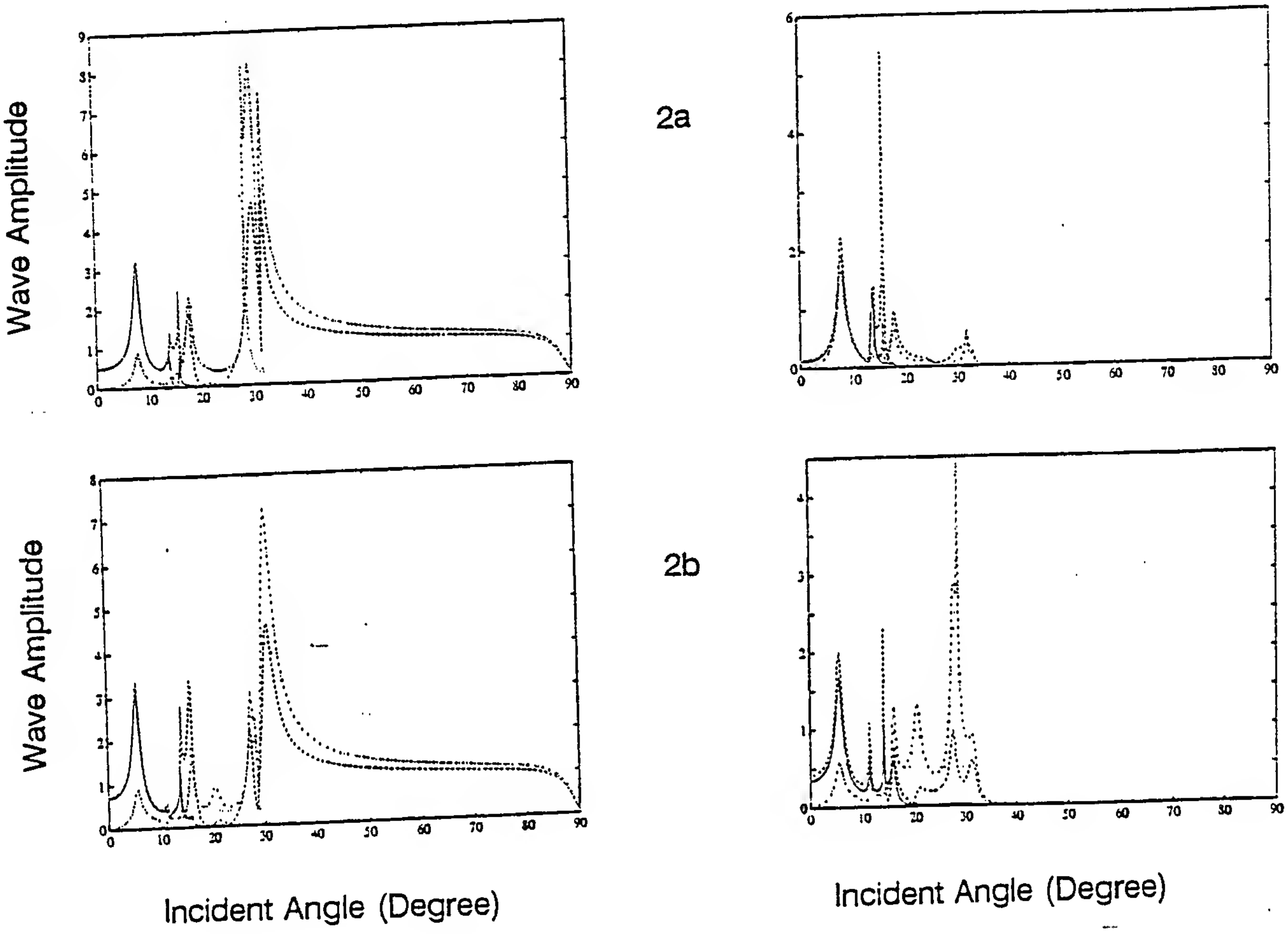


Figure 7: Wave amplitudes in top layer (left column) and bottom layer (right column) of specimens 2a (top row) and 2b (bottom row).

$7^o$ inclinations respectively. On the other hand if this angle is set at $40^o$ or greater then defects at the top layer are detected and the bottom layer defects do not have any influence on it. Similar plots for specimen 2a show that for $30^o$ incidence the top aluminum plate carries stronger waves and at an angle near $15^o$ the bottom aluminum plate has stronger wave amplitudes. Similar plots for specimen 2b failed to show any angle that generates relatively stronger waves at the bottom layer at 1 MHz. However, at another frequency it may be possible to generate relatively stronger waves in the bottom layer of this specimen.

## CONCLUDING REMARKS

The calculations presented in this paper show that it is possible to detect and image defects at individual layers minimizing the influence of other layers on the image. It can be done by generating Lamb waves of appropriate mode which produces strong excitation in a specific layer of the plate. The theoretical analysis is carried out using the Thomson-Haskell matrix method with delta matrix modification.

## REFERENCES

1. W. T. Thomson, J. Appl. Phys. 21, 89 (1950).
2. N. A. Haskell, Bull. Seism. Soc. Am. 43, 17 (1953).
3. J. W. Dunkin, Bull. Seism. Soc. Am. 55, 335 (1965).
4. B. L. N. Kennett and N. J. Kerry, Geophys. 57, 557 (1979).
5. D. L. Folds and C. D. Loggins, J. Acoust. Soc. Am. 62, 1102 (1977).

6. D. B. Bogy and S. M. Gracewski, Int. J. Solids Struct. 20, 747 (1984).

7. A. H. Nayfeh and T. W. Taylor, in Review of Progress in QNDE, Vol.7B (1988).

8. T. Kundu and A. K. Mal, Wave Motion 7, 459 (1985).

9. T. Kundu, J. Acoust. Soc. Am. 83, 18 (1988).

10. D. Levesque and L. Piche, J. Acoust. Soc. Am. 92 (1992).

11. B. Hosten and M. Castaings, in Review of Progress in QNDE (1993).

## PUBLICATIONS

The following papers have been published/prepared by T. Kundu from his research works sponsored by the AFOSR office.

1. T. Kundu and M. Blodgett, "Detection of Material Defects in Layered Solids Using Lamb Waves", Review of Progress in Quantitative Nondestructive Evaluation, Vol.13, Eds. D. O. Thompson and D. E. Chimenti, Pub. Plenum Publishing Co., QNDE Conference in Brunswick, Maine, Aug.1-6, 1993.

2. T. Kundu and B. Maxfield, "A New Technique for Measuring Rayleigh and Lamb Wave Speeds", Journal of the Acoustical Society of America, Vol.93(6), p.3066-3073, 1993.

3. M. A. Awal and T. Kundu, "V(z) Curve Synthesis Using Two Ultrasonic Transducers", ASME Journal of Applied Mechanics, submitted, 1994.

# WAVELET ANALYSIS OF ULTRASONIC SIGNALS FOR NON-DESTRUCTIVE EVALUATION OF COMPOSITES

Theresa A. Tuthill
Assistant Professor
Electrical Engineering Dept.

University of Dayton
300 College Park Dr.
Dayton, OH 45469-0226

Final Report for:
Summer Research Extension Program
Wright Laboratories

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

The University of Dayton

December 1993

# WAVELET ANALYSIS OF ULTRASONIC SIGNALS FOR NON-DESTRUCTIVE EVALUATION OF COMPOSITES

Theresa A. Tuthill
Assistant Professor
Electrical Engineering Dept.
University of Dayton

## Abstract

Wavelets are an innovative, computationally efficient method for the time-frequency analysis of ultrasound signals. This study looked at some fundamental traits associated with the wavelet transform coefficients as applied to the non-destructive evaluation of composite materials. A software package utilizing the discrete wavelet transform was developed to facilitate the study of many of these features.

Based on the Daubechies family of wavelets, an "optimum" wavelet kernel was determined for use with ultrasound scan lines. Application of this wavelet with a thresholding algorithm provided an efficient data compression technique. For use in flaw detection, the time-frequency scalogram image was enhanced using adaptive histogram equalization. However, characterization of materials based on patterns in the wavelet transform were hampered by a shift variance. And, finally, the frequency-dependent attenuation coefficient was extracted from the transform coefficients, though a large error in accuracy remained. The wavelet transform shows promise in non-destructive evaluation of materials, though further work is needed.

# WAVELET ANALYSIS OF ULTRASONIC SIGNALS FOR NON-DESTRUCTIVE EVALUATION OF COMPOSITES

Theresa A. Tuthill

## I. INTRODUCTION

The characterization and detection of defects in composite materials is integral to the evaluation of structures. Ultrasound provides an efficient, inexpensive tool for examining material composition without direct observation and is commonly used in non-destructive evaluation (NDE) applications. However, the resolution provided by such techniques often precludes detection of localized anomalies.

As an ultrasound pulse propagates through a medium, it is refracted, scattered, and attenuated, depending on the material's internal structure and properties (specifically, variations in sound speed and density). Information characterizing the material is thus included in the time signal amplitude as well as derived acoustic parameters such as the backscatter and attenuation coefficients. These acoustic parameters, however, are often dependent on frequency, and a spectral analysis is necessary.

Detection of flaws and microcracks is highly dependent on the resolution of the resulting scans which is limited by the insonifying pulse width and frequency. Flaws smaller than a wavelength will not be detected using standard amplitude intensity display techniques. Incorporation of phase information, while retaining localized spatial resolution, will improve detection.

This study examines an innovative signal processing technique for combining time-frequency analysis - the discrete wavelet transform. The motivation for examining wavelets is based on a phenomenological approach to ultrasound echoes : echo pulses look like wavelets.

## II. DISCUSSION OF THE PROBLEM

The evaluation of a material's strength and structural integrity is often difficult because damage below the surface cannot be resolved. Delaminations, microcracking, and debonding are common flaws which when left undetected can propagate with catastrophic results.

For a flawed material or inhomogeneous medium, the ultrasound signals are non-stationary, and localized changes to frequency dependent parameters would be undetectable in a spatially smeared windowed spectrum. Thus a time-frequency representation is more applicable.

An inherent problem of a time-varying spectrum, however, is founded in the uncertainty principle. The product of the time resolution and frequency resolution remains a constant. Increasing frequency resolution reduces time (or spatial) resolution and vice versa. Wavelet analysis can optimize both resolutions and offers an accurate and efficient approach for characterizing materials.

Though wavelet theory has been a burgeoning research topic in the past few years, its application to ultrasound signals warrants specific attention. An appropriate wavelet must be determined to optimize efficiency in flaw detection and characterization. Additionally, a fundamental understanding of the resulting wavelet transform patterns must be gained for evaluation purposes.

## III. THEORY

### 1) Wavelet Theory

As an alternative to windowed Fourier analysis techniques, the wavelet analysis [1] was chosen for its ability to locate time discontinuities and its invertible transform. A wavelet family is comprised of translations and scale variations of a single wavelet and can be described [ 2] by

$$h_{a,b}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right) \qquad (1)$$

where a represents the time scale parameter and b is the time shift. For a<1, the the waveform is compressed, and the frequency is correspondingly increased. The resulting family of wavelets forms a set of basis functions for the decomposition of a given time signal. That is, a one-dimensional signal can be expressed as a linear combination of the members in a given wavelet family. The contribution of each wavelet is given by it's wavelet transform coefficient. The wavelet transform of signal f(t) is just the correlation of the wavelets with the signal and is defined as

$$Wf(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\, h^*\left(\frac{t-b}{a}\right) \partial t = \left\langle f, h_{a,b} \right\rangle \qquad (2)$$

The signal can be reconstructed from its wavelet transform coefficients using the following equation

$$f(t) = \frac{1}{C} \int_{0}^{\infty} \int_{-\infty}^{\infty} Wf(a,b)\, \frac{\partial a\, \partial b}{a^2} \qquad (3)$$

where C is a constant dependent on the wavelet.

The choice of a specific "mother" wavelet (the original h(t)) is dependent on the application. For an exact reconstruction, the wavelet must meet the following criteria:

1) admissibility (the wavelet has zero mean)

2) orthonormal

and 3) finite energy.

Ingrid Daubechies, a reknowned mathematician, has done extensive work in determining orthonormal basis wavelets that are compactly supported [3]. Figure 1 shows a Daubechies wavelet and its corresponding Fourier Transform. Note that the wavelet is essentially a bandpass signal.

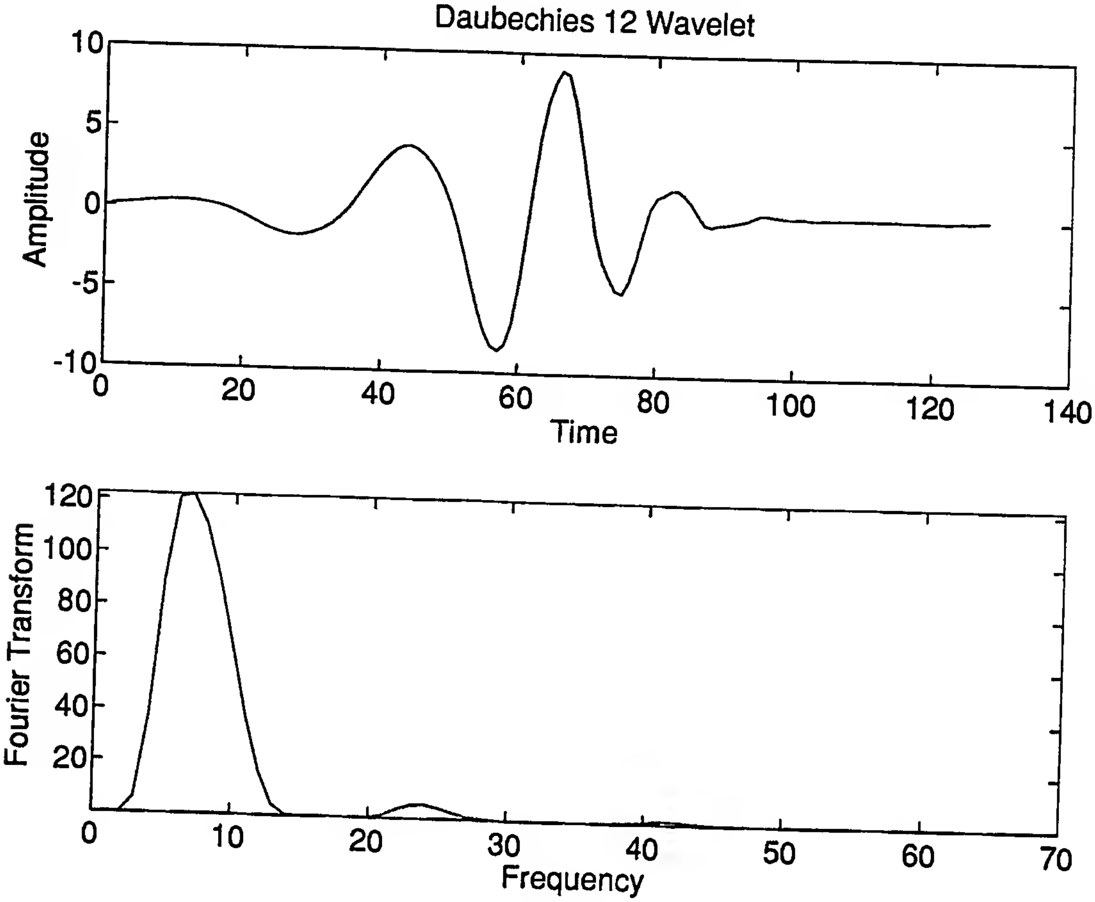


Figure 1: Daubechies 12 wavelet and its corrsponding spectra.

In the frequency domain, the wavelet transform behaves like a multi-frequency channel decomposition [4]. Each wavelet scale has a frequency bandwidth proportional to its center frequency thus providing a constant frequency-to-bandwidth ratio. The implementation of the wavelet transform serves to filter the signal through these bandpass filters.

The multi-resolution technique for determining the wavelet transform is readily implemented through digital filtering [5]. Different scale views of a signal are created by sub-sampling. Then the signal at any given scale can be separated into two frequency bands: high (corresponding to the wavelet function) and low( corresponding to the scale function). The high frequencies would represent the detail and are related to the transform coefficients. Figure 2 shows the discrete wavelet transform implemented with filter where g(n) is a high pass filter and h(n) is a low pass.

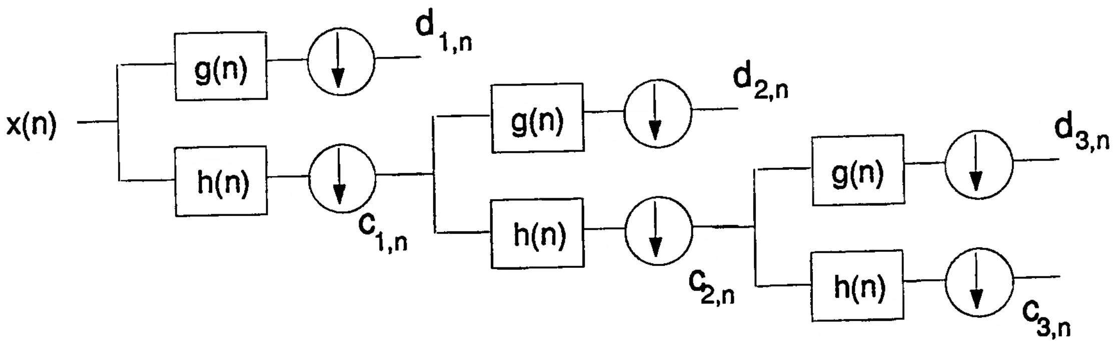


Figure 2: Implementation of discrete wavelet transform

The original signal is separated into its high and low frequency components, and both are subsampled by two. The high frequency output is then stored as the wavelet transform coefficients, dn. The low pass signal, cn, has the process repeated on it. Note that at each stage or resolution level the number of coefficients is halved. The wavelet transform preserves the number of coefficients; an N-point data signal results in N wavelet coefficients.

To provide an exact reconstruction of the signal, the above filters must meet specific criteria:

1) The filters must be quadrature mirror filters (QMF). Their Fourier transforms are thus related by the equation

$$|H(\omega)|^2 + |G(\omega)|^2 = 1 \qquad (4)$$

2) The filter impulse responses are given by the following

$$\sum h(n) = \sqrt{2} \qquad (6)$$

$$\sum h(n) = \sqrt{2} \qquad (6)$$

where the $\sqrt{\ 2}$ normalization factor is due to the decimation factor in the filter.

3) The filters must be "regular such that the filtering iterations converge.

Extensive research has been done in the area of determining filters that meet the above criteria. Using maximally-flat low pass filters, Daubechies computed wavelet filter coefficients for various filter lengths. Because the corresponding wavelets have continuous derivatives and are compactly supported, they have become the "standard" in many wavelet applications.

Displaying the wavelet transform requires a two-dimensional, time-frequency plot. By using the squared modulus of the transform coefficients, the energy of the signal can be represented on a gray scale plot called a scalogram. Time or distance is shown on the horizontal axis and increasing frequency is shown on the vertical axis going from bottom to top. Because the number of wavelet coefficients decreases with decreasing frequency, the lateral resolution decreases at each successive level, and more pixels are assigned to a single coefficient. The highest resolution occurs for the highest frequency and is displayed on the top row. Any localized or transient phenomena should appear in this row.

2) Acoustic Parameters

The loss of energy as an acoustic pulse propagates through a medium is referred to as attenuation. Attenuation encompasses longitudinal and shear scattering form inhomogeneities and conversion of acoustic energy to heat by absorption mechanisms. The resulting decay in signal strength can be fitted to an exponential of the form

$$S(x) = s_o \, e^{-2Ax} \tag{7}$$

where x is depth, A is the frequency-dependent attenuation coefficient, and $s_o$ is a constant representing the backscattered signal strength. A standard procedure for estimating the attenuation coefficient at the center frequency of the pulse is to curve fie the decay of the envelope of a scanline. However, this technique doe not provide frequency information.

## IV. METHODOLOGY

1) Facilities

The research was performed in the University of Dayton's Ultrasound Lab which was established in 1991. The lab was designed for the implementation of research in the areas of both biomedical tissue characterization and non-destructive evaluation of materials. A single ultrasound scanning station is used for data collection, signal processing, and data analysis. Samples are placed in a gallon water tank and can be scanned at various frequencies ranging from 1 to 10 MHz using Panametrics transducers. A computer controlled, X-Z positioner, with high precision stepper motors (with 1/100" increments), is used to move the transducer across the scanning plane. The output from a Panametrics pulser/receiver is digitized via an HP45402 digitizing oscilloscope (400 Msamples/sec, 2K sample storage). The data is transmitted over the IEEE bus to the SUN-IPC SPARCstation (16" color monitor, 24Mbytes RAM, 2.2Gbyte hard drive) for storage and processing.

The MATLAB software package (Mathworks, Inc., Natick, MA) with the Signal Processing Toolbox is the primary programming environment. MATLAB allows users to incorporate their own C-code for analysis and then display results using a variety of visualization techniques. An Ethernet local area network also provides communication and file sharing with other computers for remote processing.

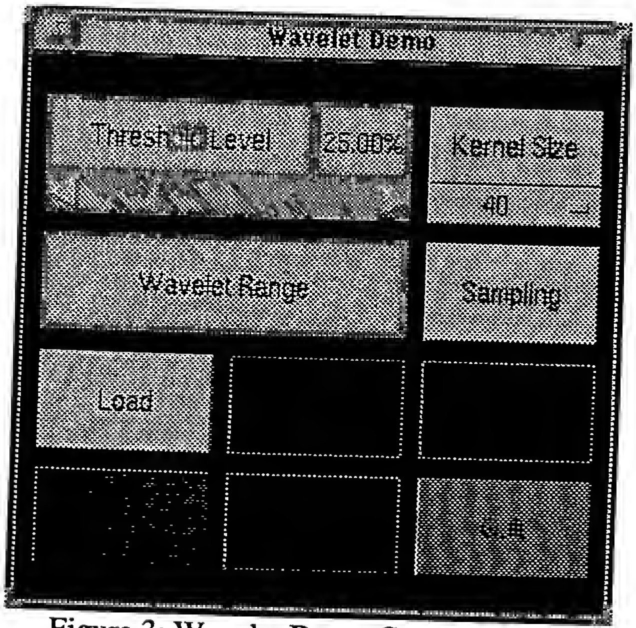
2) Implementation of Wavelet Transform

A fast discrete wavelet transform was implemented within the MATLAB environment. A suitable algorithm, found in Numerical Recipes in C (Second edition), was coded and interfaced so that it could be called directly from the MATLAB command prompt. The code was adapted and expanded to include the Daubechies kernels with 4 to 20 coefficients.

The wavelet transform can be thought of as a transformation matrix acting on a column vector of data. The transformation matrix is sparse for large values of N (data vector length), which reduces the number of calculations needed to compute the wavelet transform. The transform can also be thought of as two related convolutions, in which only half of the values are kept from each convolution. The first convolution serves as a smoothing filter, while the second as a high pass filter. The two convolutions are interleaved after the first pass, then the data is reordered and a second pass is made on the N/2 samples from the first convolution. This method is similar to the algorithm for the FFT. The process is repeated until only 2 values are generated from the first convolution. The resulting vector contains the wavelet transform coefficients and can be thought of as blocks of amplitudes for a given frequency band, with each band twice as big as the previous band.

3) Wavelet Demo Software Package

A control software package was written to provide an easy-to-use, analytical tool for wavelet analysis. First, scan line signals stored in MATLAB data files are loaded into the program. User interface control buttons then let the user change various parameters related to the wavelet transform. Figure



Threshold Level 25.00%
Kernel Size 40

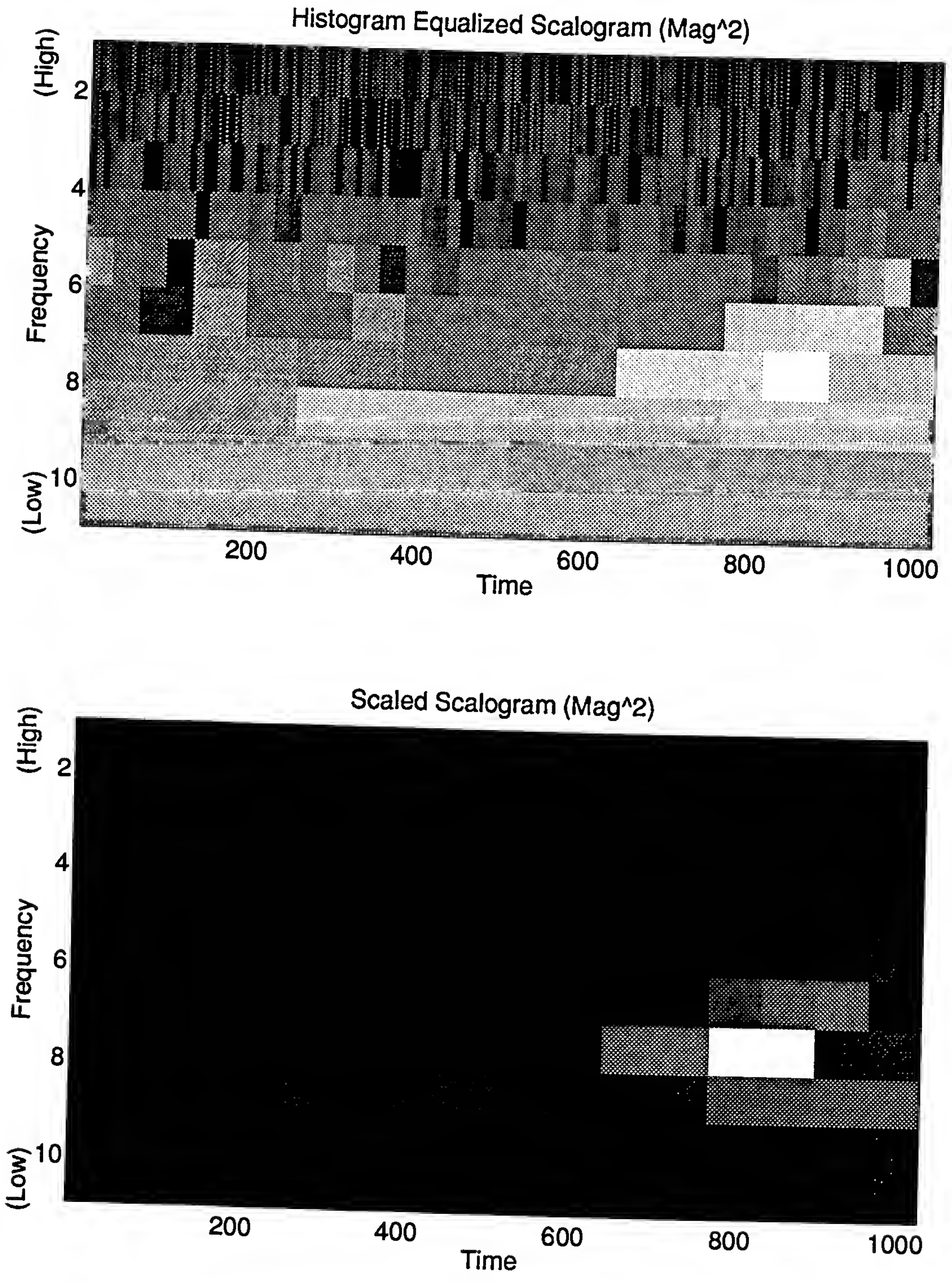
Figure 3: Wavelet Demo Control panel.

3 shows the interface display with controls for the threshold level (for truncating coefficients), kernel size (number of coefficients for the filters), wavelet range (for zooming in on a specific range), and the number of points desired for downsampling. The program then computes and displays four basic graphs: the original waveform, the wavelet transform, the reconstructed waveform, and the two-dimensional scalogram. The wavelets transform is displayed by combining the coefficients from each band into a single vector. The highest frequency band is to the left using points N/2 to N, the second highest band is in points N/4 to N/2, and so on down to the lowest frequency band contained in the first two points.

Through the course of this research project, the scalogram was analyzed for patterns relating to the material. The original display mapped the coefficient values to the full 255 gray levels. However, because the high frequency coefficients were much lower in magnitude, they were often "washed out," precluding any identification of high frequency transient effects. As a result, a histogram equalization was performed within each frequency band. Figure 4 shows a scalogram from a typical waveform before and after the equalization process.

4) Composite Materials

Ultrasound scans were taken of a variety of composite materials. To analyze a "damaged" composite, scans were made of a graphite-epoxy, quasi-isotropic, 3/32" thick sample with a 4.05 ft-lb impact. A scan of a non-damaged region is referred to as "QND010," and a scan of the damaged region is "QND150."



Histogram Equalized Scalogram (Mag^2)

Scaled Scalogram (Mag^2)

Figure 4: Scalograms before and after histogram equalization.

## V. RESULTS

Ultrasound scan lines from the QND010, QND150, and a 32 ply uni-weave composite sample were formed using a 3.5 MHz, 2" focused transducer and then processed for the following analyses.

### 1) Optimum Wavelet

Throughout this study, various Daubechies wavelets were employed. The different wavelets are denoted by their kernel size (or length when considering filter coefficients) ranging from 4 to 40. Inspection of the corresponding waveforms shows that the size 20 kernel most resembled the insonifying pulse in smoothness and in length. This observation was confirmed in the data compression analysis discussed below. To illustrate, Figure 5 shows the 4-point kernel wavelet transform for a scanline (1000 points using a 5 MHz transducer and sampling at 200 MHz) taken of the 32 ply composite sample. In the reconstructed signal, only a portion of the coefficients are used, and the result is a jagged waveform. Figure 6 shows the wavelet transform of the exact same signal, but now using the 20-point kernel. Again, only a portion of the coefficients were used to reconstruct the signal, though here the resulting waveform is much smoother and resembles the original. Despite the vast difference in the resulting waveforms, the mean-square-errors (MSE) between the original and the reconstruction are close.

It should be noted that despite the jaggedness of the 4-point kernel wavelet, when all the coefficients are employed, an exact reconstruction is formed. This was tested with all the kernel sizes and confirms the assertion that the Daubechies wavelets are invertible.

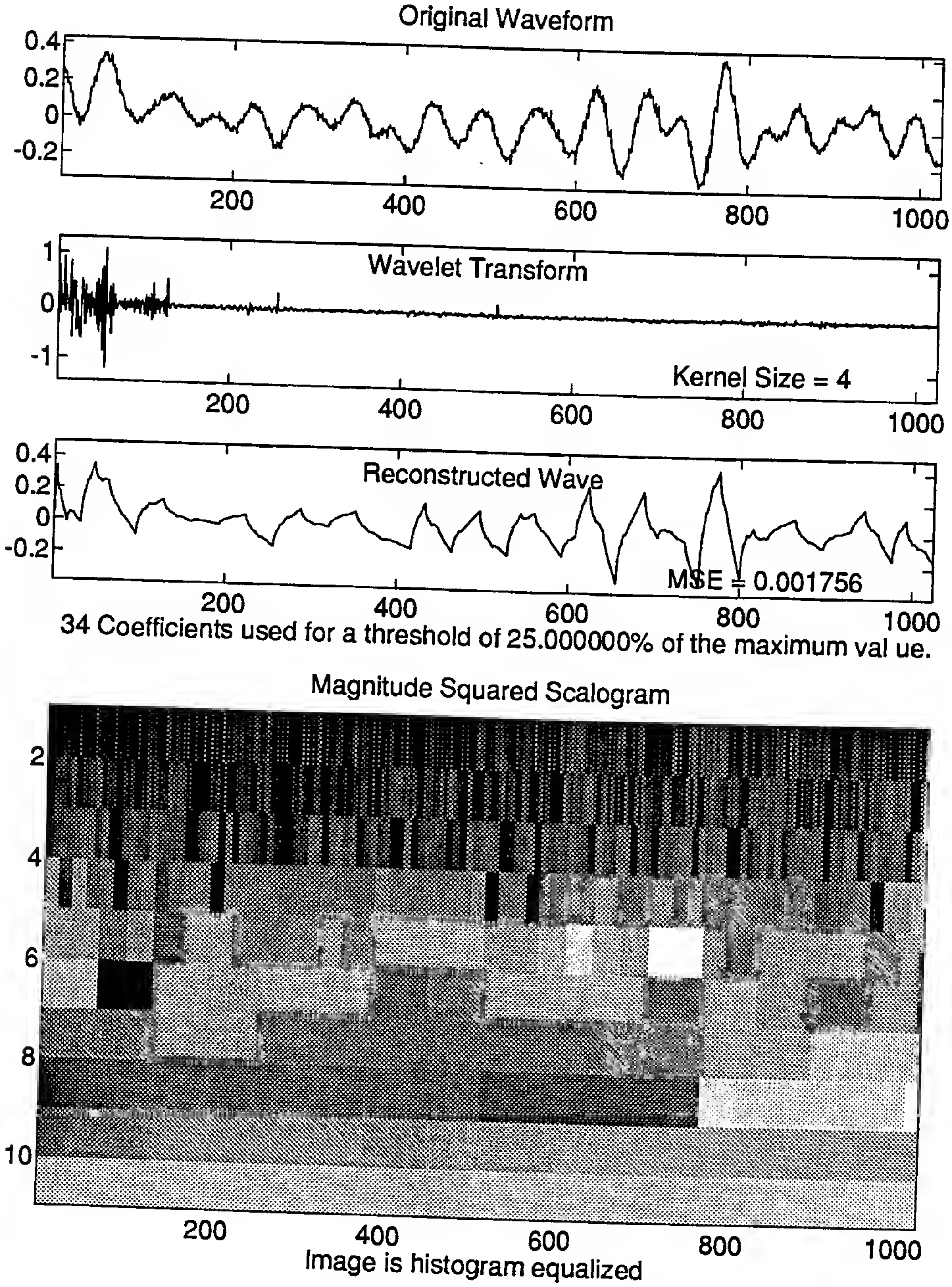


Figure 5: Wavelet transform using 4-point kernel.

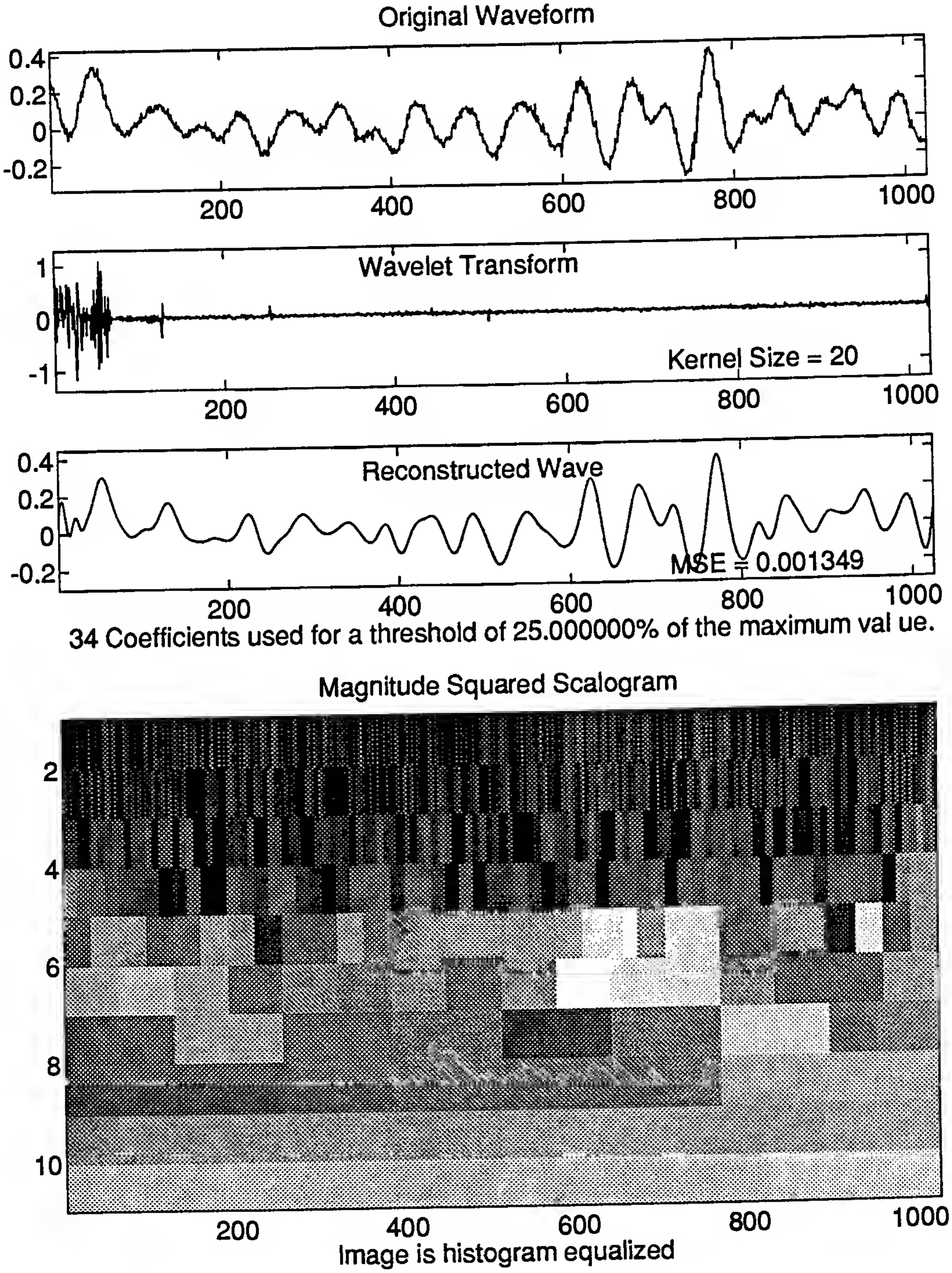


Figure 6: Wavelet transform using 20-point kernel.

2) Data Compression

The wavelet transform is suitable for data compression because it is orthogonal and has a finite number of basis functions. Since the original signal's energy is compressed into a fewer number of coefficients in the transform space, some of the coefficients can be eliminated from the reconstruction without loss of information. The analysis of data compression focused on two techniques: i) truncation of transform coefficients to a given number of points and ii) thresholding the transform values and keeping only those coefficients that were above a specified level. The metric for evaluating the compression techniques is the mean square error between the reconstructed signal and the original input signal.

i) Truncation - Since the wavelet transform orders the coefficients into frequency bands going from lowest to highest frequency, truncation is similar to downsampling the data. When transform coefficients are zeroed, the high frequencies are removed and a smoother or low pass filtered signal will be reconstructed. Large errors occur if the signal is originally not smooth or contains high frequencies. A range of truncations were performed on the 512-point wavelet transform of the scanlines from the QND composite sampled at 100 MHz . Figure 7 shows the results for truncating the number of coefficients down to 64 or a compression of 8 to 1. The corresponding mean-square-error (MSE) of 8.5 provides a reasonable reconstruction. Further analyses showed that a MSE value of 10 is a reasonable cut off point for determining a good compression scheme. However, further truncation down to 32 coefficients (a compression ratio of 16 to 1) produced an MSE of 253 (Figure 8) and a poor reconstruction.

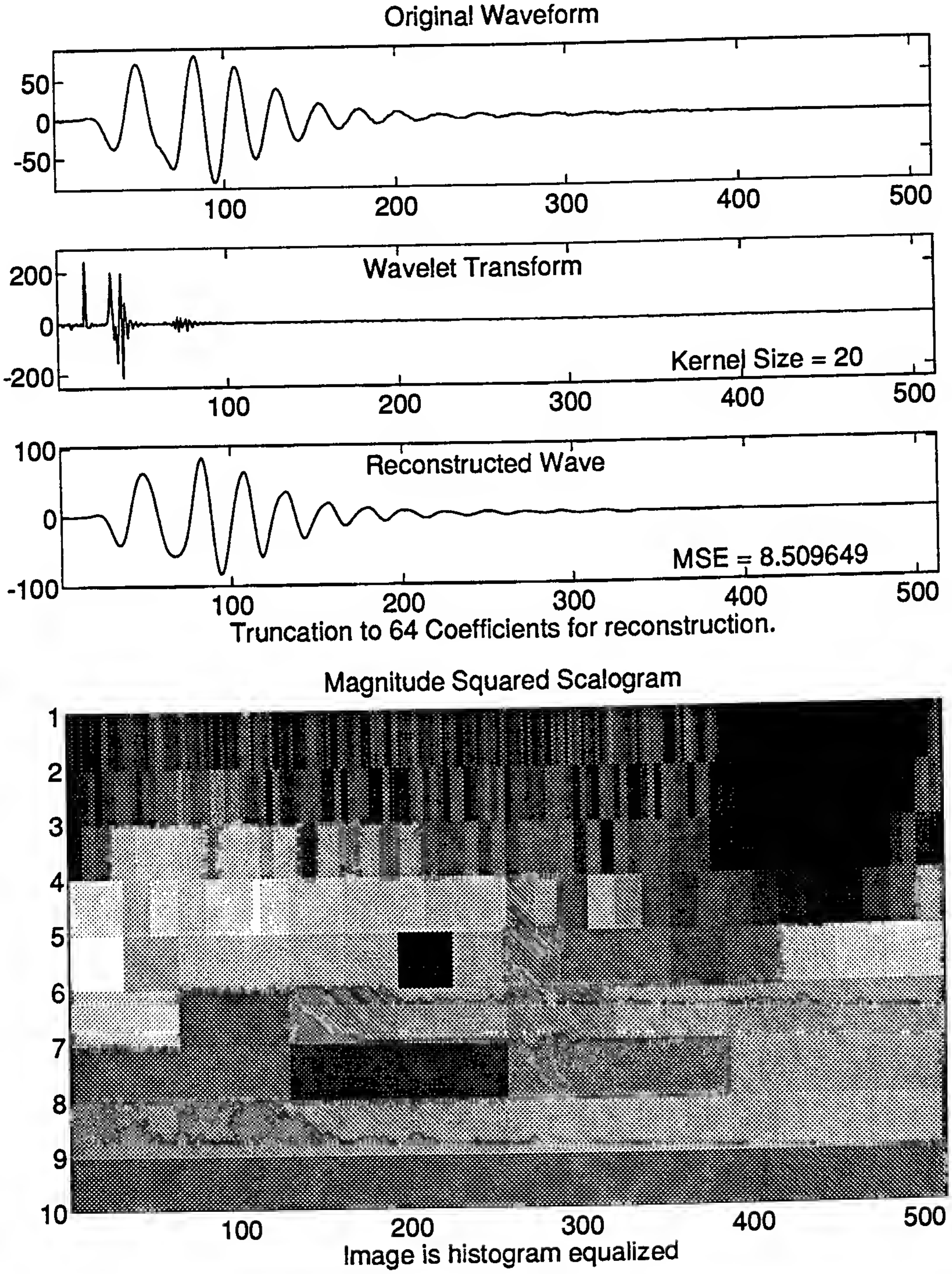


Figure 7: Signal reconstruction after truncating wavelet transform to 64 coefficients.

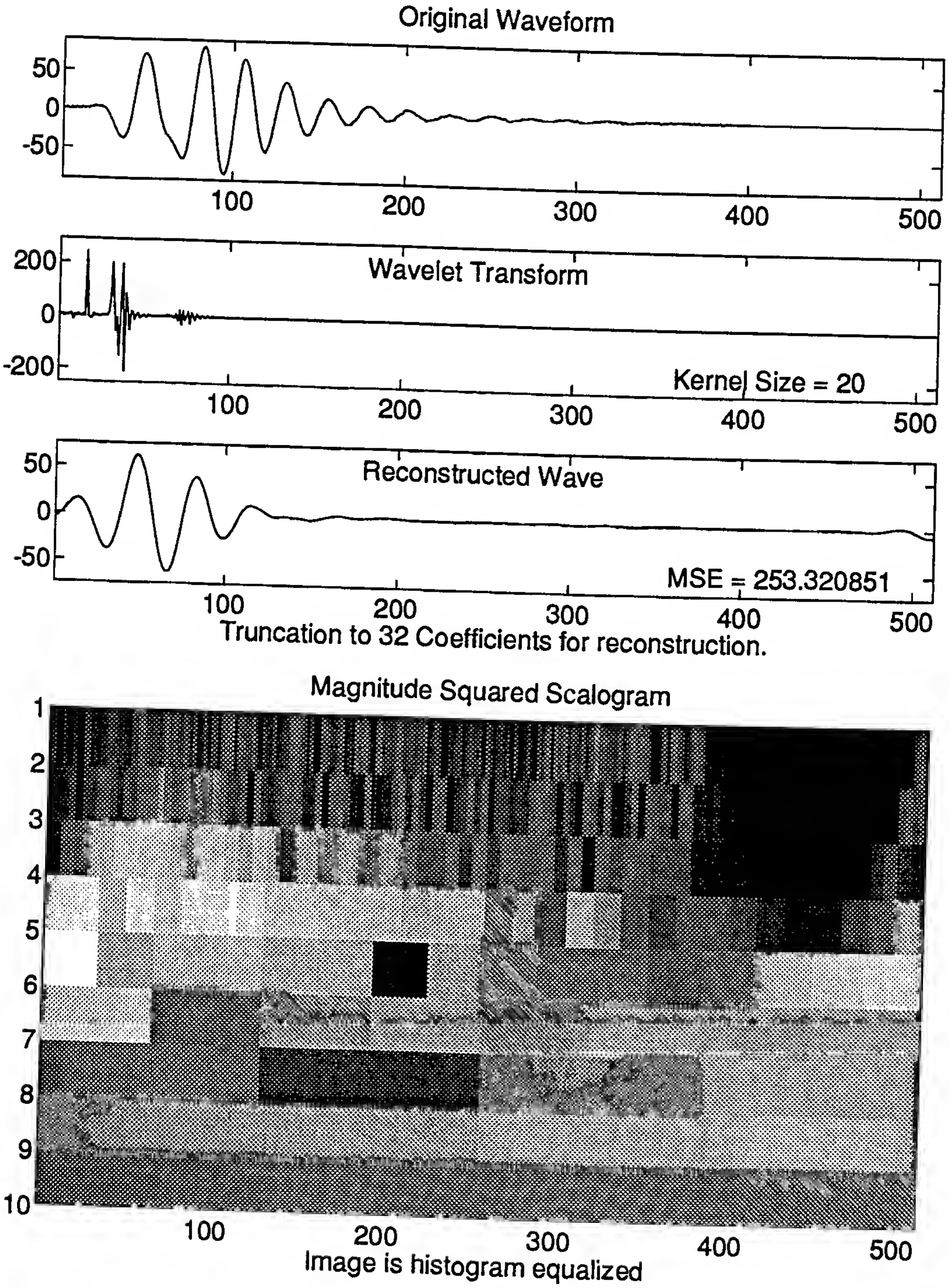


Figure 8: Signal reconstruction after truncating wavelet transform to 32 coefficients.

This technique did lead to interesting application for removing high frequency spikes from an input signal. Since the wavelet transform has both frequency and time information a spike in the input signal could be removed by attenuating or zeroing out the proper coefficients in the transform that are near the time period of the spike. This could be thought of as a notch filter localized in time around the spike.

ii) Thresholding - Data compression can also be obtained by storing only those coefficients with values above a certain threshold level. With this technique, however, it is necessary to also store the corresponding position of the coefficient in the resulting transform vector. A program was written in C which performs the thresholding of the data. The program searches for the absolute maximum value then computes the threshold level and zeros all transform coefficients below that threshold. This does add some overhead, since the value of the transform and the location must be saved. However, this method still provides better compression of the signal and lower MSE for the reconstructed signal.

Table 1 - Thresholding errors for constant threshold levels

| Kernel Size | # Coeffs | Threshold level | Error (MSE) |
|---|---|---|---|
| 4 | 11 | 33% | 58.5 |
| 12 | 8 | 33% | 41.9 |
| 16 | 9 | 33% | 82.7 |
| 20 | 6 | 33% | 43.2 |
| 24 | 4 | 33% | 76.4 |
| 34 | 9 | 33% | 32.2 |
| 40 | 6 | 33% | 64.4 |

Table 2 - Thresholding errors for constant number of coefficients.

| Kernel Size | # Coeffs | Threshold level | Error (MSE) |
|---|---|---|---|
| 4 | 6 | 46% | 139.5 |
| 12 | 6 | 42% | 83.4 |
| 16 | 6 | 25% | 50.3 |
| 20 | 6 | 33% | 43.2 |
| 24 | 6 | 27% | 32.2 |
| 34 | 6 | 49% | 84.5 |
| 40 | 6 | 33% | 64.4 |

The size of the kernel or wavelet does effect the efficiency of the thresholding. Table 1 shows that for the wavelet transform coefficients of a scanline from the QND sample using a constant threshold level, the number of resulting coefficients first decreases as the kernel size increases, but then starts to increase again. As expected, the MSE varies inversely with the number of coefficients. It does appear, however, that an "optimum" wavelet kernel size exists around 20 to 30. In support of this theory, the errors were computed for the various kernel sizes using a set number of coefficients. Table 2 shows that there is a minima for the MSE in the 20 to 30 range. For an ultrasound scan line it is expected that the optimum kernel size would be comparable to the wavelength, and for the preceding example of a 3.5 MHz pulse sampled at 100 MHz, the number of sample points for a wavelength is in fact 28.

A direct comparison between the truncation and thresholding techniques was done on the QND sample using the 20-point kernel and varying the threshold until the desired number of coefficients was achieved. Figure 9 shows a threshold level of 0.8 percent of the maximum value is necessary for retaining 64 coefficients. This corresponds to storing 128 points (position must be included) or a compression ratio of 4 to 1 The error was only 0.20 in reconstructing the signal. By increasing the threshold until only 32 coefficients are used (Figure 10), the MSE increases to only 1.19. Examination of the reconstructed waveforms shows that the signal can be represented by very few coefficients and still provide a good reconstruction of the signal. Higher threshold error levels could achieve compression ratios of 50 to 1 with tolerable MSE values. This technique provides better results and higher ratios than the truncation method, however there is a higher amount of overhead in determining, which values to keep .

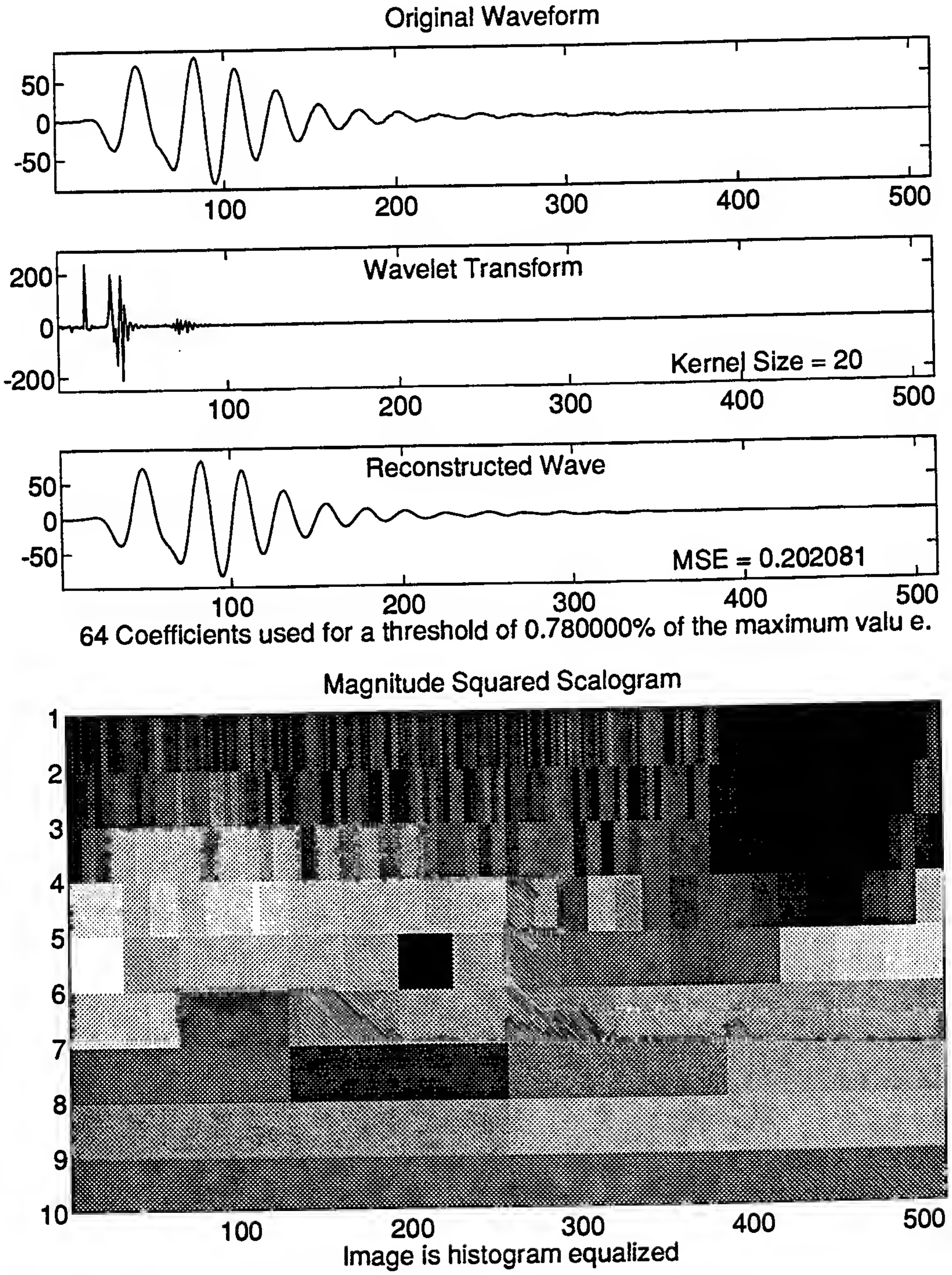


Figure 9: Signal reconstruction after thresholding wavelet transform to 64 coefficients.

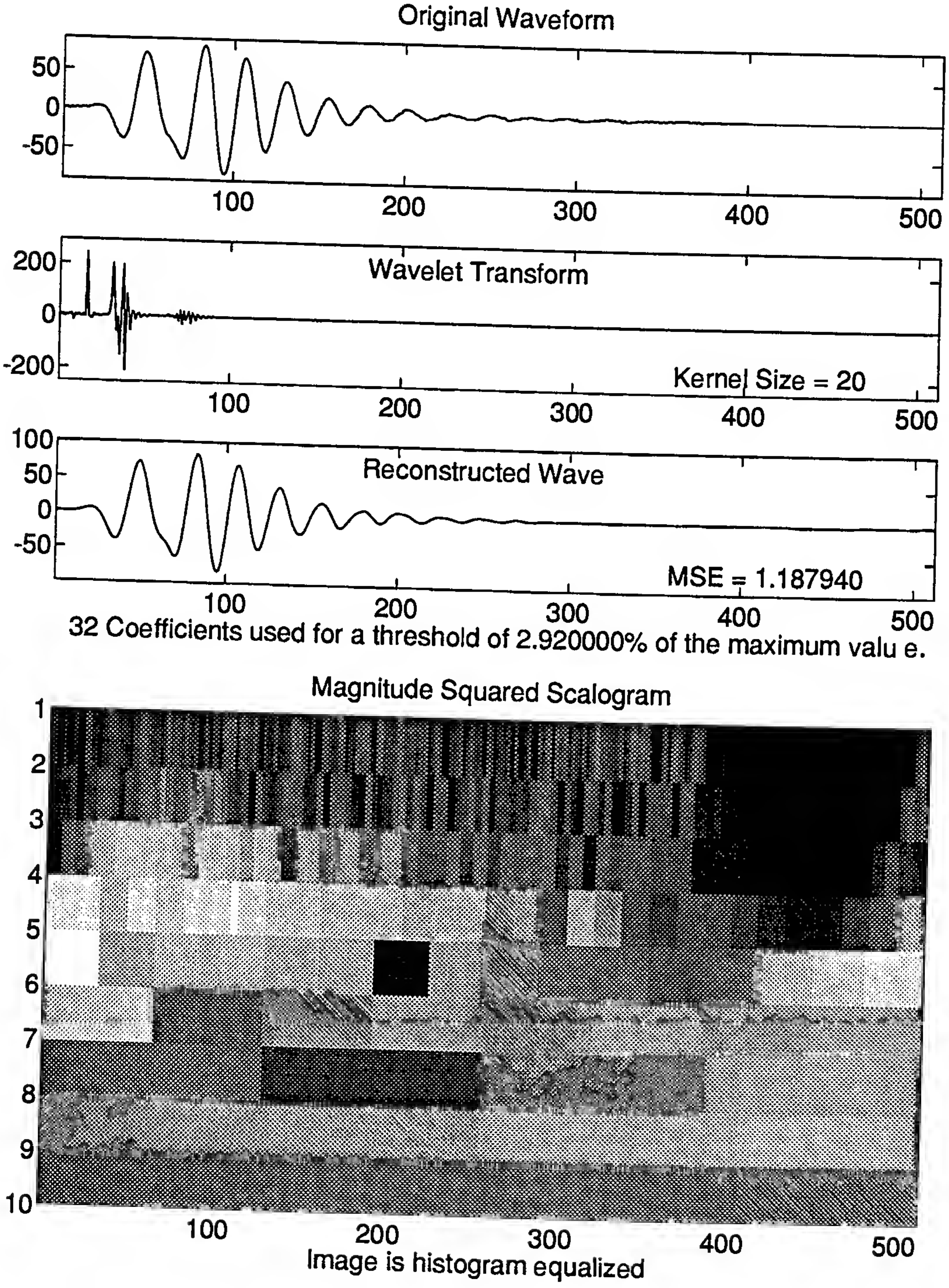


Figure 10 : Signal reconstruction after thresholding wavelet transform to 32 coefficients.

3) Effect of Sampling Rate on Data Compression

The data analyzed in the previous section was over-sampled, so the effect of the sampling rate on the compression ratios was examined. The input signals were down-sampled by various factors of 2, and then reprocessed to determine new compression statistics. Only the thresholding method was evaluated in this section since it had the best performance. Figure 11 shows a 512-point, 3.5 center frequency scanline sampled at 100MHz. A threshold level of 33% produces a reasonable reconstruction with a MSE of 43. Figures 12 and 13 show this same waveform down-sampled by a factor of 2 and 4, respectively. In both cases, the number of coefficients needed for a reconstruction error near 40 was approximately the same as for the original signal. Note that for each decimation by two, the highest frequency band in the scalogram is eliminated. And, while the resulting scalograms are not identical for the lower frequency bands, the patterns are similar. This study showed that the sampling rate will not effect compression rations, as long as the signal is sampled well above the Nyquist..

4) Shift Variance

The most significant problem with applying wavelet analysis to ultrasound signals is that the wavelet transform is shift-variant. The resulting transform coefficients depend on the position of the input signal in the time domain. Even small shifts in the time domain can lead to significant changes in the wavelet domain. Figures 14 and 15 show the wavelet transform coefficients for shifted versions of a 128 point sequence of the QND sample. Even for this narrow range of shifts (-3 to +4), there is a noticeable difference in the transform, especially in the lower frequency bands. The coefficients with the largest magnitude are fairly similar between 1-point shifts, though the amplitude goes from positive to negative through the entire sequence of shifts.

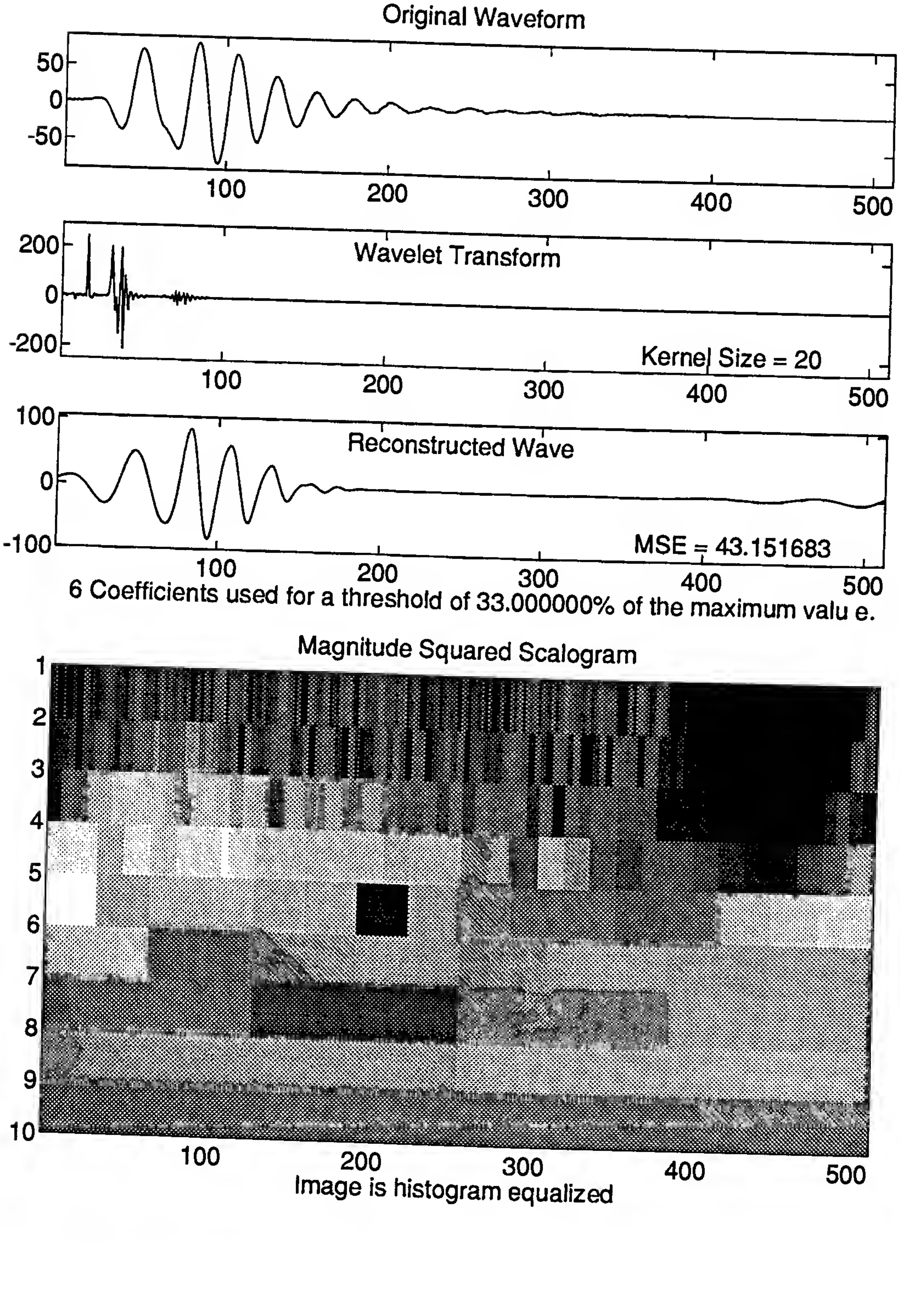


Figure 11: Wavelet transform of signal sampled at 100 MHz.

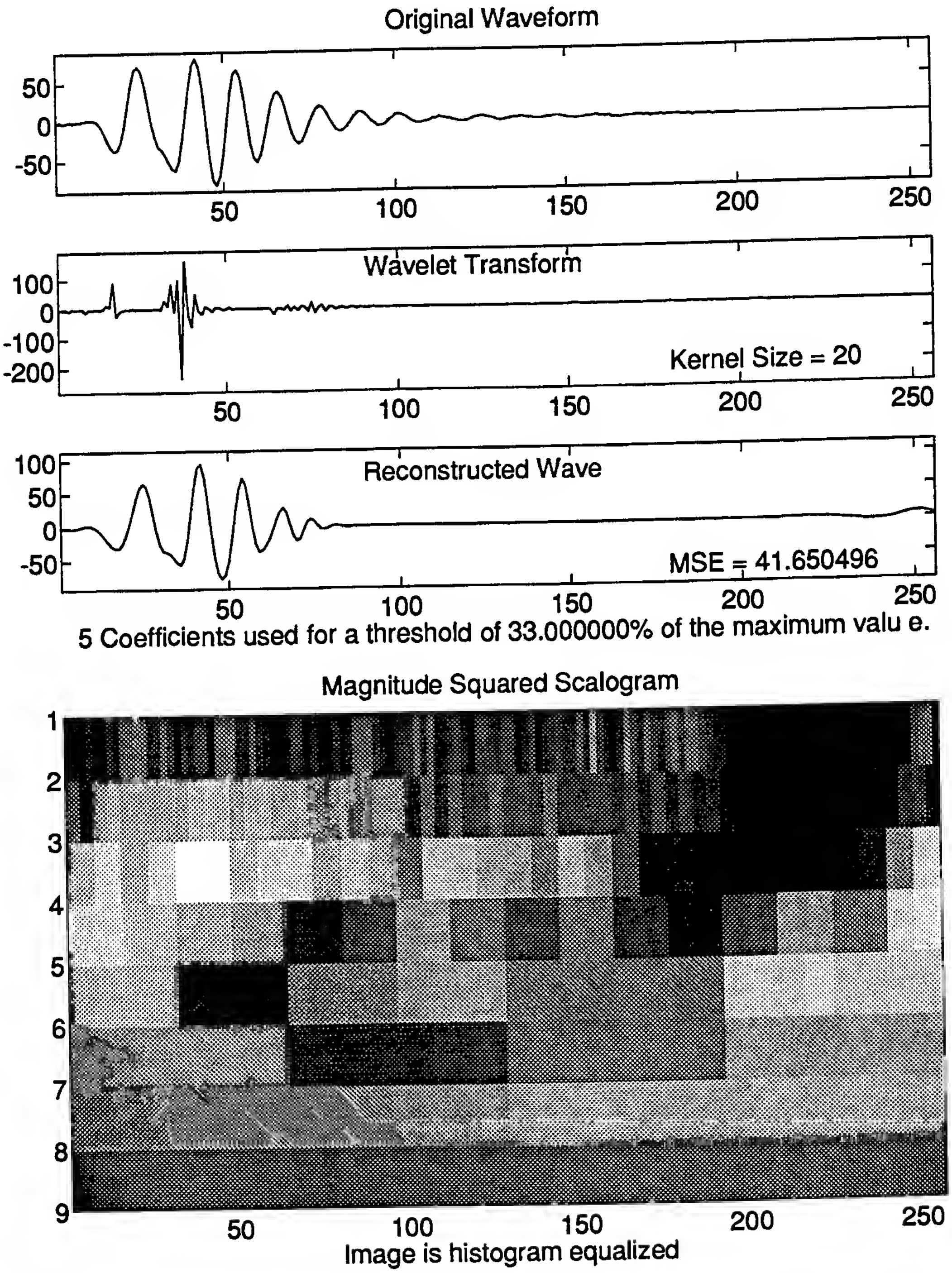


Figure 12: Wavelet transform of signal (same as in Fig. 11) sampled at 50 MHz.

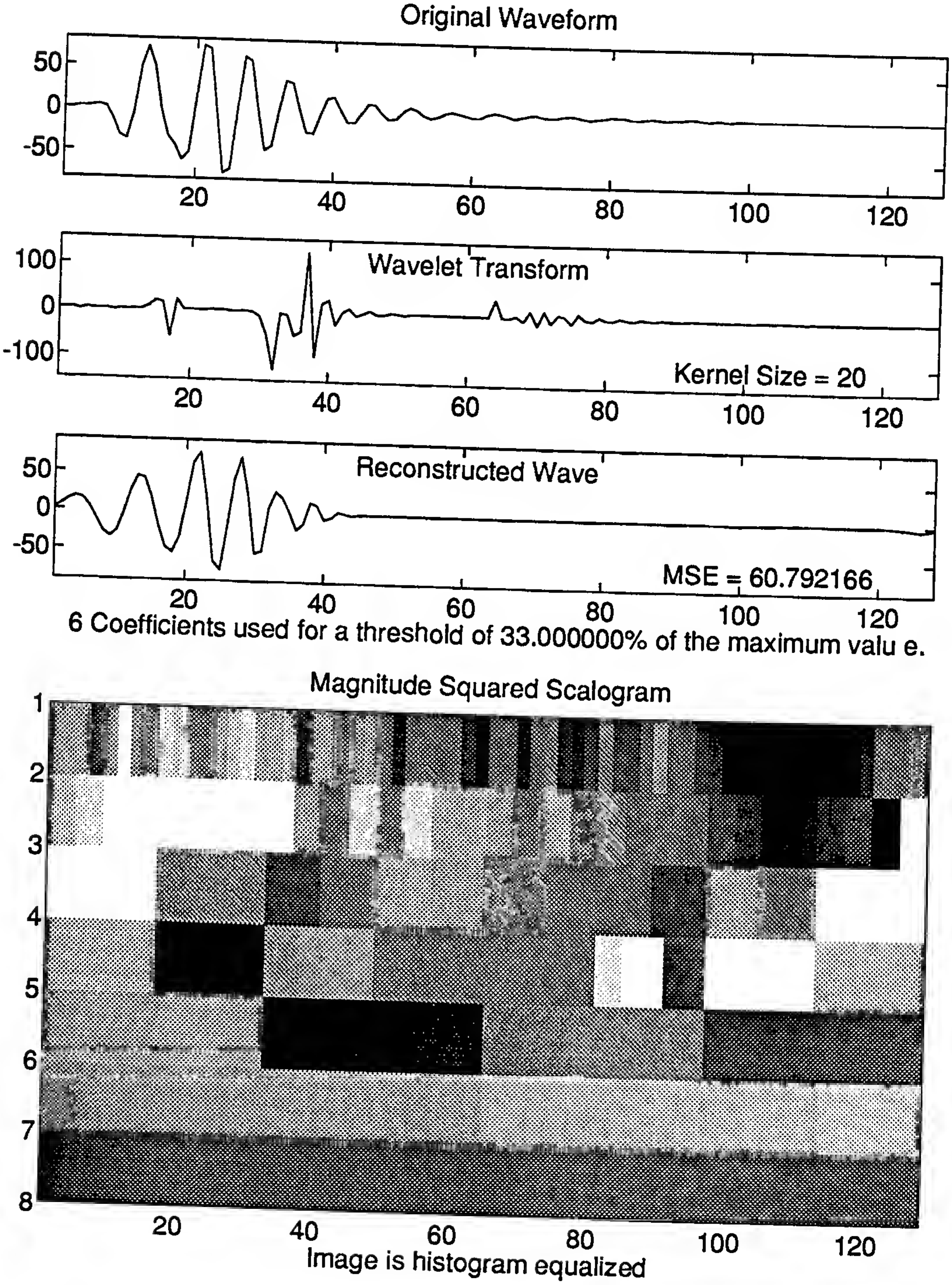
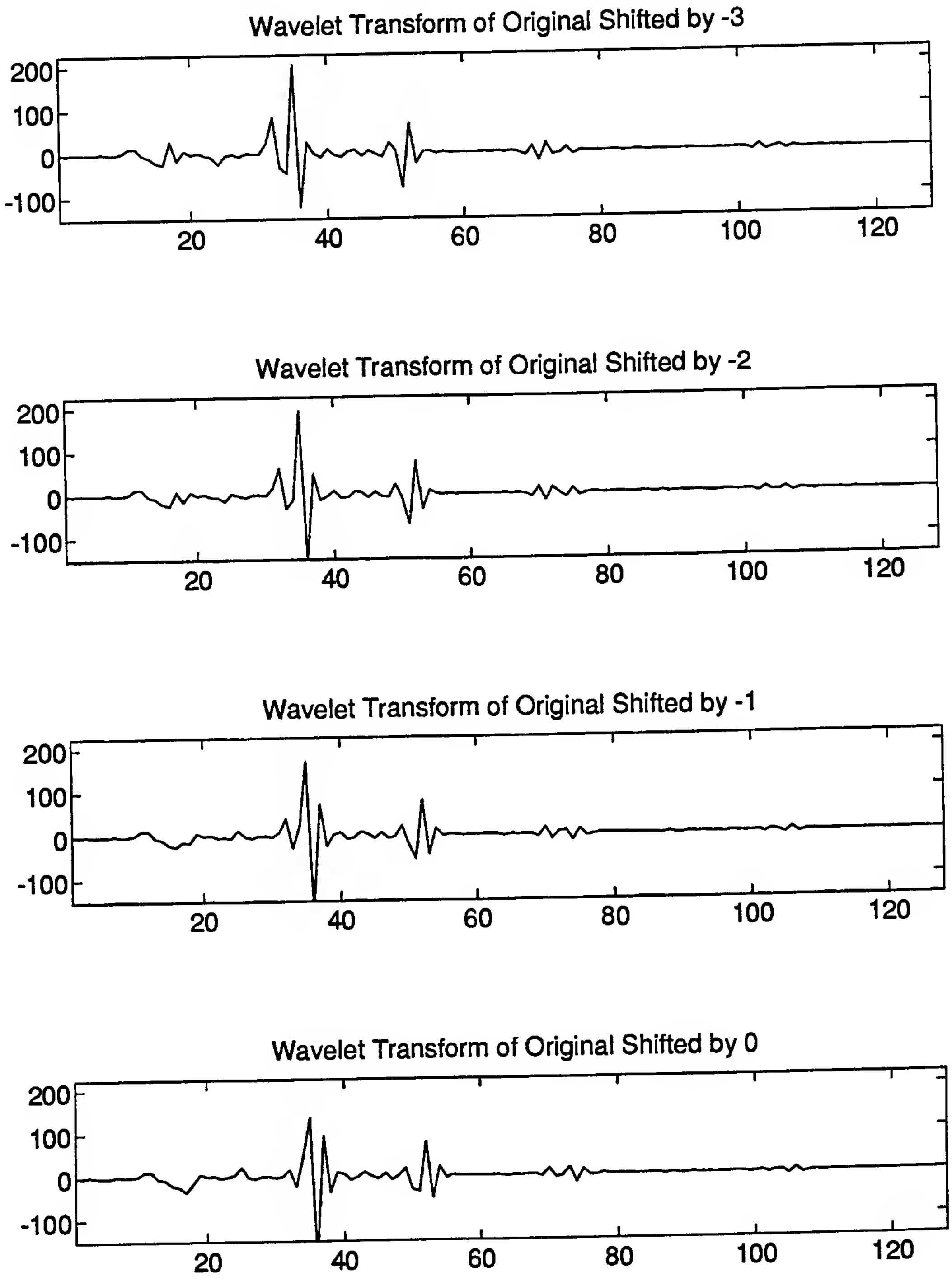


Figure 13: Wavelet transform of signal (same as in Fig. 11) sampled at 25 MHz.



Wavelet Transform of Original Shifted by -3

Wavelet Transform of Original Shifted by -2

Wavelet Transform of Original Shifted by -1

Wavelet Transform of Original Shifted by 0

Figure 14: Wavelet transform of signal shifted by negative increments.

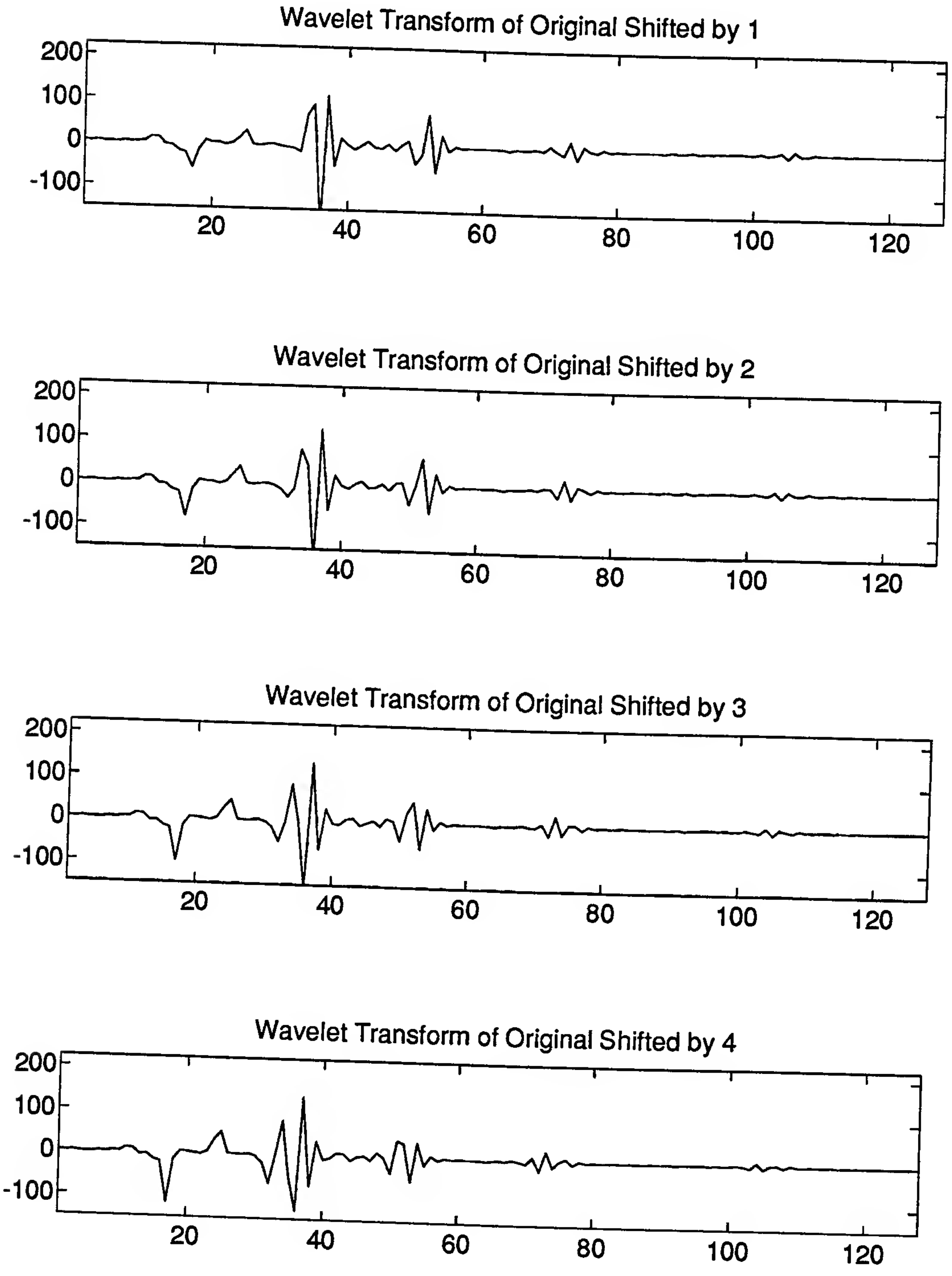
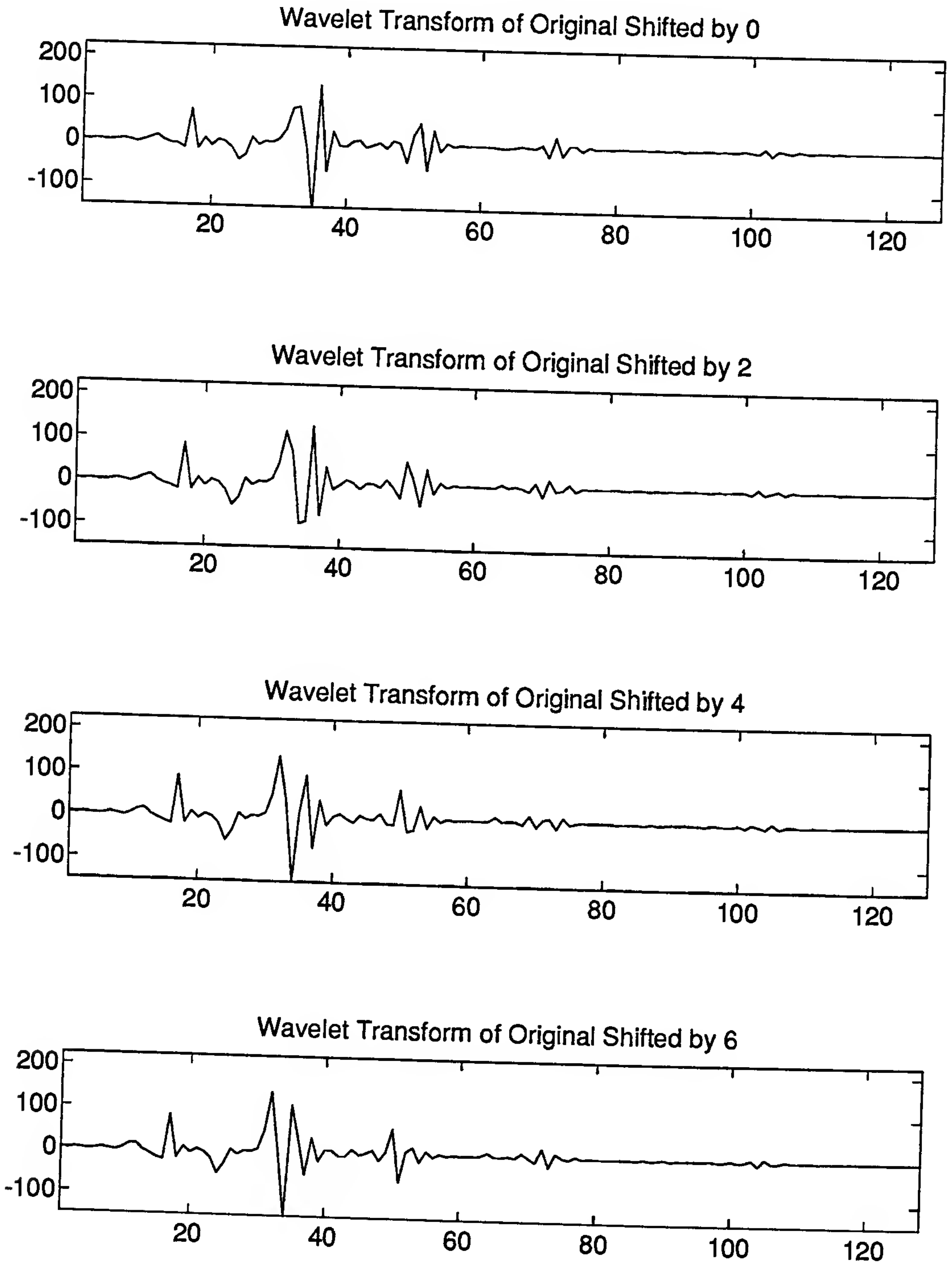


Figure 15: Wavelet transform of signal shifted by positive increments.

In computing the discrete wavelet transform, a decimation by two is performed after each filtering stage. It is expected then, that within a specific frequency band, shifts by a corresponding multiple of two would produce a simple shift within that band. For example, a shift by 4 would correspond to a shift by 2 in the second highest frequency band (points 32 through 63 in a 128 point transform) and a shift by 1 in the highest band (points 32 through 63). The wavelet transforms of a 128-point sequence (using a 20 point kernel size) are shown for shifts of various powers of two in Figures 16 and 17. Close examination of the higher frequency bands shows that the coefficients are shifted, but still show a small variation. It is concluded that the high pass filter applied before decimation in determining the wavelet coefficients is the main cause of this minor variation. The coefficients in the lower bands, however, appear uncorrelated.

Shifts in the time domain are an inherent problem with ultrasound scans since jitter in the pulser/receiver can cause delay shifts for adjacent scan lines.. The vastly different wavelet transform coefficients can preclude a reliable characterization based on wavelets.. To try and correct this problem the signals were cross correlated and then shifted so that they had maximum correlation.

The cross-correlation was computed for adjacent signals in the QND B-scan, then the maximum value was found and a shift computed from how far the maximum value was from the center of the correlation vector. This proved to be effective if the signals were similar, however if the signals were repetitive (i.e. a strong back echo), the cross-correlation would correlate different peaks and the shifted signals would be even worse than the original signals. Improvements were made by correlating a small section of the signal, but the amount of signal used is dependent on knowledge of what the signals look like and how much shift is in the signals.

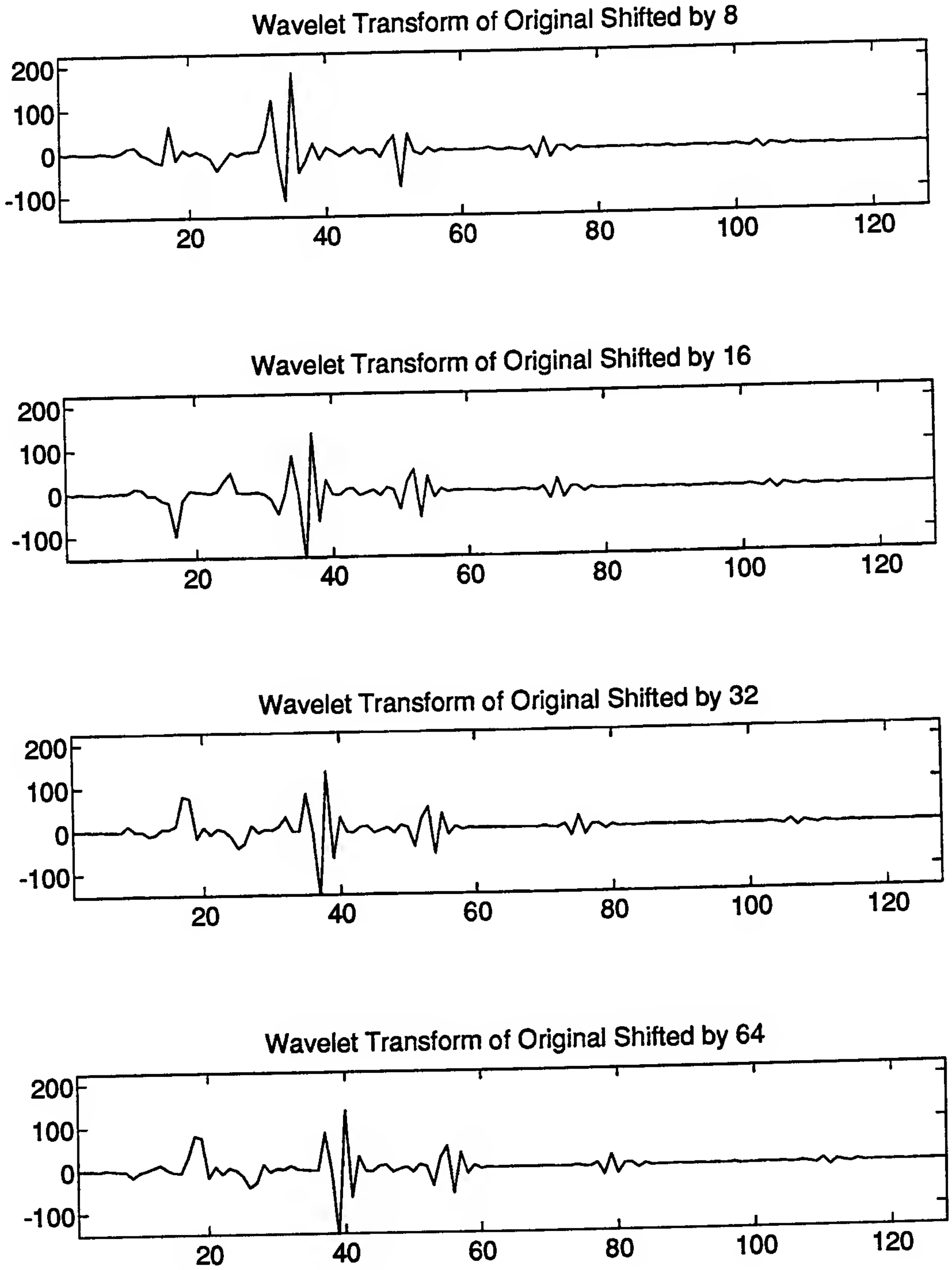

Wavelet Transform of Original Shifted by 0

Wavelet Transform of Original Shifted by 2

Wavelet Transform of Original Shifted by 4

Wavelet Transform of Original Shifted by 6

Figure 16: Wavelet transform of signal shifted by powers of two.


Wavelet Transform of Original Shifted by 8

Wavelet Transform of Original Shifted by 16

Wavelet Transform of Original Shifted by 32

Wavelet Transform of Original Shifted by 64

Figure 17: Wavelet transform of signal shifted by powers of two.

5) Signal Attenuation

The integrity of a composite material can often determined by the attenuation of the ultrasound signal as it passes through the material. The attenuation coefficient, however, is frequency dependent, and a frequency domain analysis is required. Either FFTs are used to compare attenuated power spectra, or the power outputs from a bank of filters are compared. This latter technique was adapted for the wavelet transform frequency bands. To test the accuracy of this technique, a metal stepper block with uniform step sizes was scanned, and the wavelet analysis was applied to the front and back echoes. The average front and back scan power was found for each step and each individual frequency band. The log of the "back echo power" to the "front echo power" was then computed and plotted versus the depth of the material for each frequency band. The slope, which is directly related to the attenuation, was determined by applying a linear curve fit. To insure zero attenuation for zero depth, the curve was forced through the origin. Figure 18 shows the results using a 5 MHz transducer for filter bank 5 (corresponding to a low frequency band near 5 MHz). The curve fits for bands much higher or lower were not as accurate due to the low amplitude strength of the transform coefficients in these band (Figure 19). This experiment was repeated using a 10MHz transducer. Again, the best curve fits were found for those filter banks nearest the 10MHz center frequency (in this example banks 3, 4, and 5). The slope for the three most accurate banks is shown in Figure 20 plotted versus frequency. Note that for both transducer data sets, as the frequency increases (corresponding to smaller filter bank numbers), the attenuation increases. The accuracy (or lack thereof) of employing wavelet coefficients for extracting attenuation values is demonstrated by the difference in slope values for a given filter bank.

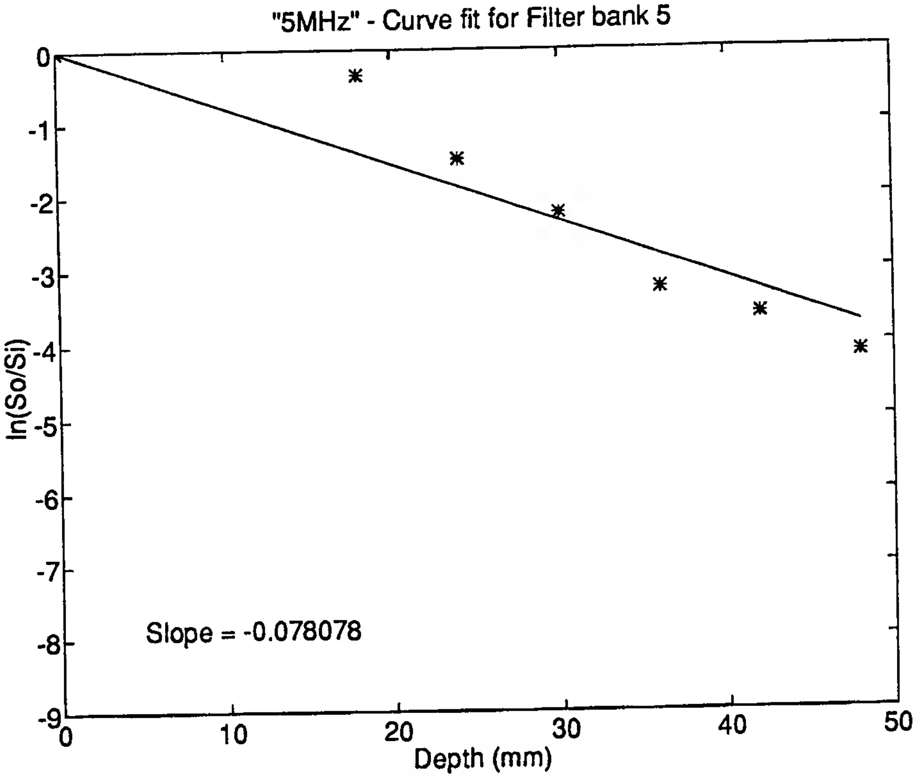


Figure 18: Plot of attenuation versus depth for the 5th filter bank (with a 5 MHz) transducer.

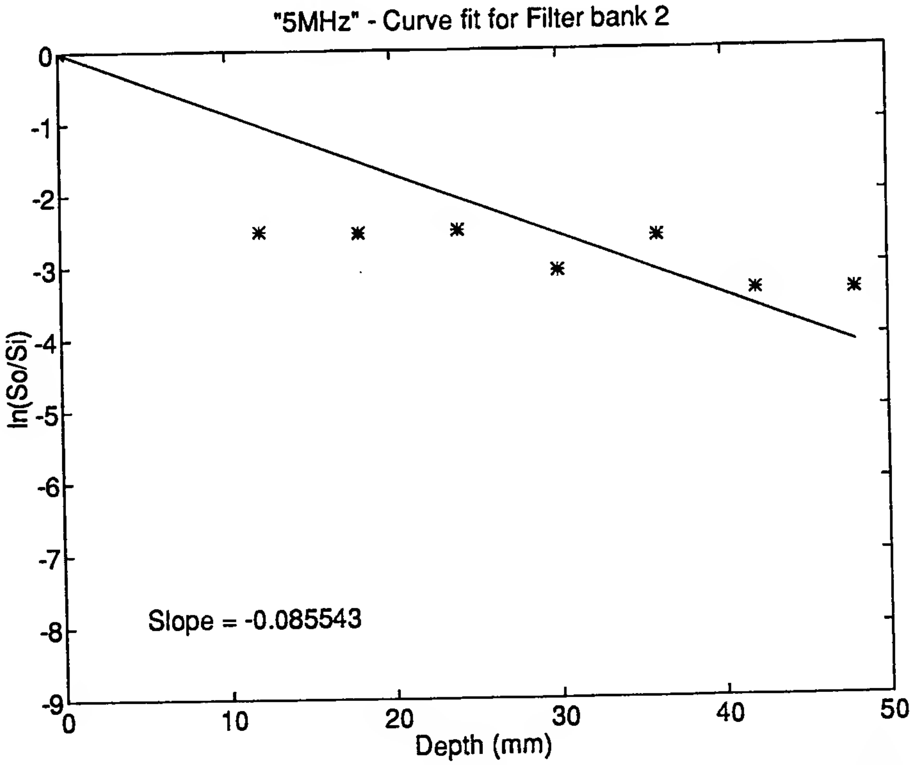


Figure 19: Plot of attenuation versus depth for the 2nd filter bank (with a 5 MHz) transducer.

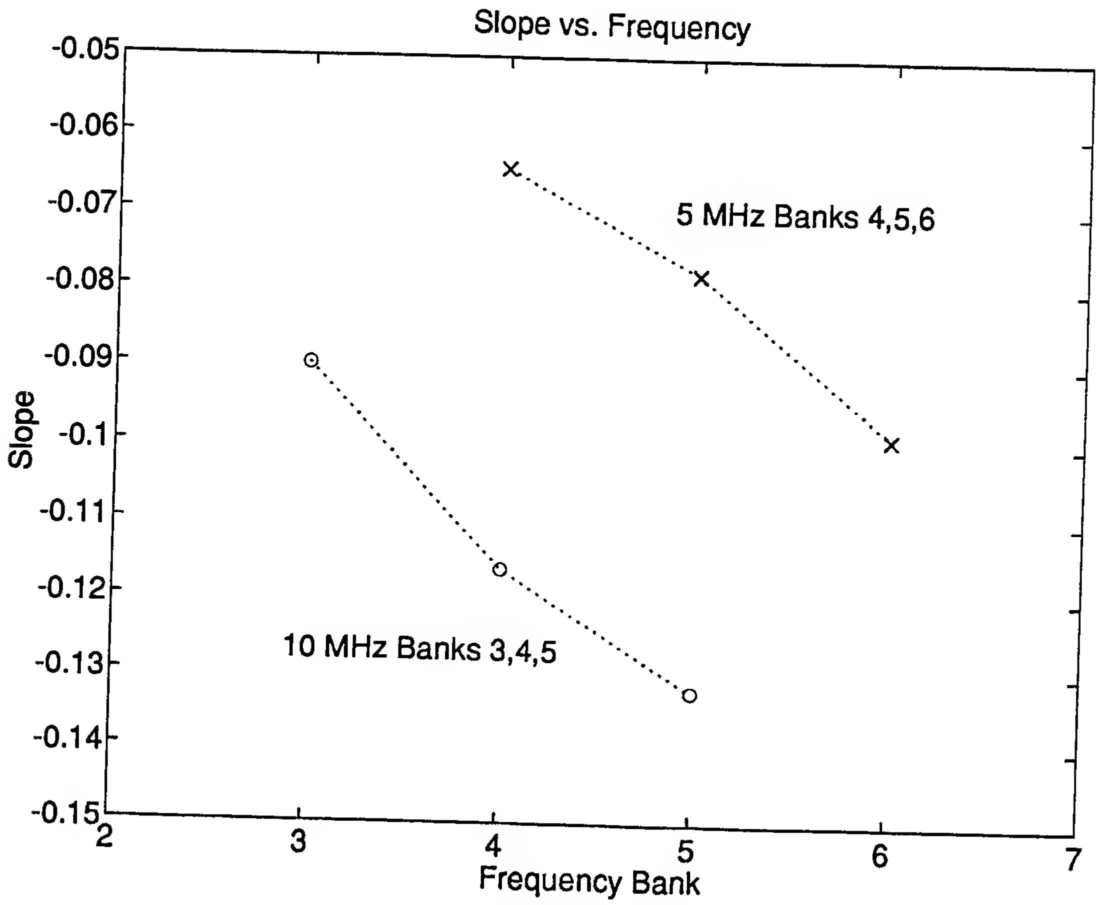


Figure 20: Attenuation versus frequency for both 5 and 10 MHz transducers.

The cause of the separation between the attenuation lines for the 5 MHz data and the 10 MHz is unclear. While these results are promising, additional research is needed to determine if these values are in fact accurate for given applications and to determine the range of spatial information needed for such precision.

## VI. CONCLUSION

The wavelet transform does hold potential as a computationally fast, efficient analysis tool for evaluating composite materials. A Daubechies wavelet, with a kernel size comparable to the length of the digitized insonifying pulse, produces the best results for ultrasound scans. By thresholding the resulting wavelet transform coefficients, the scan lines can be compressed with minimal error in the reconstruction. Depending on the application and corresponding resolution, compression ratios of 50 to 1 can be achieved.

Wavelet transform scalograms can depict transient flaws providing a histogram equalization is applied. Characterization of debondings and cracks based on the scalogram patterns still requires more research. Part of the problem in material characterization based on wavelet coefficients is the inherent non-linear shift variance of the wavelet transform. Techniques must still be devised to account for unknown shift variations. Finally, the localized attenuation coefficient can be obtained form the wavelet coefficients, but the error associated may preclude accurate assessment..

While none of the analysis results demonstrate a remarkable breakthrough, the discrete wavelet transform has provided a unified basis for a combined analysis. This research has shown that wavelets can be useful in the non-destructive evaluation of composites and may be instrumental in the future automation of such processes.

## ACKNOWLEDGMENTS

The author is indebted to Edward Meyer, a Ph.D. candidate in the Electrical Engineering Dept., who developed and coded the Wavelet Demo Software Package and performed most of the analysis simulations for this project. The author would also like to thank Dr. Tom Moran, Dr. Prasanna Karpur, and Brian Frock from Wright Labs for their suggestions and insights.

## VII. REFERENCES

[1] Chui, C.K. An Introduction to Wavelets (Academic Press, San Diego, 1992).

[2] Rioul, O. and Vetterli, M., "Wavelets and signal processing," IEEE Signal Processing Magazine, Oct. 1991, pp. 14-38.

[3] Daubechies, I., "Orthonormal bases of compactly supported wavelets," Comm. Pure Appl. Math., V. 41, pp. 909-996, 1988.

[4] Mallat, S.G., "Multifrequency channel decompositions of images and wavelet models," IEEE Trans. Acoust., Speech, Signal Processing, V. 37, pp. 2091-2110, 1989.

[5] Press, W.H., Teukolshy, S.A., Vettering, W.T., and Flannery, B.P., Numerical Recipes in C (Cambridge University Press: New York, NY, 1992)

# STOCHASTIC MODELING OF MBE GROWTH OF COMPOUND SEMICONDUCTORS

**Ramasubramanian Venkatasubramanian, Assistant Professor**

**Department of Electrical & Computer Engineering**
**University of Nevada, Las Vegas**
**Las Vegas, NV 89154**

**Final Report for:**
**Research Initiation Program**
**Wright Laboratory, (WL/MLPO)**

**Sponsored by:**
**Air Force Office of Scientific Research**
**Bolling Air Force Base, Washington, D.C.**
**and**
**University of Nevada, Las Vegas**

**December 1993**

# STOCHASTIC MODELING OF MBE GROWTH OF COMPOUND SEMICONDUCTORS

**Ramasubramanian Venkatasubramanian, Assistant Professor**

**Department of Electrical & Computer Engineering**

**University of Nevada, Las Vegas**

**Las Vegas, NV 89154**

**Abstract**

MBE growth of compound semiconductors and in-situ doping were studied using the stochastic modeling approach. Three specific problems related to MBE growth of compound semiconductors were studied: $GaAs$ growth kinetics, surface ordering in $GaAlAs$ and doping of semiconductors. The results of each of these projects were compared with experimental results and the agreement was good. Thus, the present study validates the use of stochastic modeling for the MBE growth kinetic studies. Details of results of each project and its comparison with experiments are presented under the discussion of each project.

# STOCHASTIC MODELING OF MBE GROWTH OF COMPOUND SEMICONDUCTORS

**Ramasubramanian Venkatasubramanian, Assistant Professor**

**Department of Electrical & Computer Engineering**

**University of Nevada, Las Vegas**

**Las Vegas, NV 89154**

## 1 Introduction

High speed, high frequency and low noise electronic device are currently being fabricated by molecular beam epitaxy (MBE) and studied for potential applications in information processing, signal processing and communication. Reproducible and controlled growth of these devices by MBE is possible only if the growth and doping mechanisms and their relation to growth parameters and their influence on the elecrtonic, optical and transport properties are well understood. The theme of this project is to study the MBE growth mechanisms of compound semiconductor and in-situ doping using the theoretical tool developed by the author called "*Stochastic Modeling*".

This project addresses theoretically three issues of growth: *GaAs* growth kinetics, surface ordering kinetics of *GaAlAs* and in-situ doping of semiconductors. This report is organized as follows. The *GaAs* growth kinetic study and its results and discussion are presented in section 2. The surface ordering kinetics study of *GaAlAs* and its results and discussion are presented in section 3. The doping kinetics study of semiconductors and its results and discussion are presented in section 4. Finally, conclusions are presented in section 5.

# 2 GaAs Growth Kinetics

## 2.1 Background

The performance of opto-electronic devices is degraded by interfacial roughness in heterostructures fabricated with III-V semiconductor compounds. The origin of the interfacial roughness is the surface roughness of the layer on which another hetero-layer is grown. There have been many experimental [1-9] and theoretical [9-15] attempts to understand the origin of the surface kinetic processes in the MBE growth of *GaAs* (100). Specifically, the surface roughening kinetics during MBE growth of *GaAs* (100) has been studied by monitoring the reflection high energy electron diffraction (RHEED) intensity [1]. Similar studies have also been carried out for *Ge* (100) [16]. In the case of *Ge* study [16], a kinetic surface roughening temperature above which a smooth surface remains smooth, was observed. It was concluded that the surface roughening is a result of competition between surface roughening processes such as adsorption and the surface smoothening process such as surface migration to stable sites. In the case of *GaAs* study [1], a transition temperature was observed above and below which the surface is rougher. This transition temperature was observed to be flux ratio and temperature dependent. The kinetics of surface roughening in this case was explained in terms of competition between the surface roughening processes such as adsorption and evaporation and the surface smoothening process such as the migration of atoms to energetically stable sites.

In this work, the stochastic model of MBE growth developed by the author [17-19] is employed to study the surface roughening kinetics in *GaAs* (100). In section 2.2, a discussion on the stochastic model is presented. In section 2.3, results of the study of surface roughening kinetics of (100) *GaAs* are presented and compared with that of the experimental work of Ref.[1]. The physical mechanism which describes the theoretical and experimental observations is also presented. Summary of this study is presented in section 2.4.

## 2.2 The Stochastic Model

In the hierarchy of growth simulation models, the stochastic model belongs in the category of macroscopic models. This model assumes a rigid lattice gas with nearest neighbor pair interactions. The stochastic model describes the time evolution kinetic equations of MBE growth in terms of the primary macrovariables, the concentration of atoms in the $n^{th}$ layer, C(n), and the atom-vacancy bond density in the $n^{th}$ layer, Q(n) and is given by Eqs 8a and 8b of Ref.[17]. For details of the model, the reader is directed to Ref.[17-19]. The main assumptions of the model are: (i) solid-on-solid (SOS) restriction (ii) random distribution approximation of the surface atomic configurations weighted by the energy of the configuration (iii)Arrhenius type rate equations for the surface kinetic processes (iv) exclusion of anti-site defects (v) exclusion of effects of surface reconstruction on the surface kinetic processes.

The time evolution of the macrovariables is described in terms of the rates of such kinetic processes as adsorption, evaporation and intra- and inter-layer migrations of $Ga$ and $As$ atoms. The adsorption process is allowed at all sites where the SOS restriction holds, (i.e.), the surface covalent bonds corresponding to the site from the layer below are satisfied. If the atoms arrive at non-SOS sites, then the atoms are allowed to migrate rapidly in their weakly bound physisorbed state until they find a proper site.

For the adsorption process of $As$, the species in the molecular beam is assumed to be diatomic $As_2$ which is equivalent to cracked arsenic. The stochastic model reported in Ref.[17] is suitable for monoatomic molecular species. As diatomic $As_2$ is used in this model, the terms corresponding to the adsorption process in Eqs 8a and 8b in Ref.[17] need to be modified as follows. The incorporation of two $As$ atoms in the nearest neighboring sites satisfying the SOS condition in the $2n + 1^{th}$ layer requires four $Ga$ atoms be available as nearest neighbors in the $2n^{th}$ layer as shown in Figure 1. The probability that there exists a surface $Ga$

adatom pair in the $2n^{th}$ layer, $P_1$, is given by:

$$P_1 = \frac{\tilde{N}_{GaGa}(2n)}{2C_{Ga}(2n)} \tag{1}$$

where $\tilde{N}_{GaGa}(2n)$ is the $Ga - Ga$ second nearest neighbor bond density. Three nearest neighbor pairs are required to form the surface $Ga$ adatom arrangement shown in Figure 1. Therefore, the probability that there exists surface atomic arrangements as shown in Figure 1, $P_s$, is:

$$P_s = {P_1}^3 \tag{2}$$

The concentration of $Ga$ atoms which have nearest neighbor arrangement as shown in Figure 1, $C_s(2n)$, is given by:

$$C_s(2n) = C_{Ga}(2n)\, P_s \tag{3}$$

Thus, the concentration of sites available for $As_2$ incorporation in the $2n+1^{th}$ layer, given by Eq 4 of Ref.[17] modifies to:

$$\left[\frac{\tilde{N}_{bb}(2n-1)}{2} - C_a(2n)\right] \longrightarrow [C_s(2n) - C_{As}(2n+1)] \tag{4}$$

Since the $As$ atoms are incorporating on the nearest neighbor sites as pairs, the time evolution of the atom-bond macrovariable, $Q_{As}(2n+1)$, needs to be modified as follows. For the model reported in Ref.[17], the coordination number for any surface site is 4 when considering the incorporation process monoatomic molecular beam species. Whereas, for the present study, the effective coordination number for a site is only 3 as one of the nearest neighbor site will be simultaneously occupied by the accompanying $As$ atom. This change needs to be made in Eqs. 8a and 8b. It is noted that the terms corresponding to the intra- and interlayer surface migrations need not be modified as the changes in the $As$ molecular species do not influence rates of these processes directly.

The evaporation and surface migration processes are described by Arrhenius type rate equations such as:

$$R = R_o e^{-\frac{E_{act}+nK_{aa}}{kT}} \tag{5}$$

where R is the rate and $R_o$ is the frequency factor of the process in $\sec^{-1}$. $E_{act}$ is the activation energy of the process for an isolated terrace adatom, $K_{aa}$ is the second nearest neighbor pair interaction energy in the (100) plane and $n$ is the number of nearest neighbors of the atom under consideration. Thus, the activation energy term appearing in the argument of the exponential is coverage dependent with $n$ equal to zero and four for low and high coverages, respectively. In general, the $E_{act}$ for surface migration is smaller than that of evaporation. In this study, the $E_{act}$ for interlayer and intralayer migrations are assumed to be equal. The surface migration of an atom to non-SOS site is not allowed.

The model parameters for the present study of the MBE growth of $GaAs$ were obtained from the literature and the MBE growth parameters of Ref.[1]. The atom pair interaction energies for the second nearest neighbors $Ga - Ga$ and $As - As$ were obtained as 0.25 eV and 0.325 eV, respectively, based on the data reported in Ref.[20]. The frequency factors for the evaporation and migration processes were chosen as 1.0 x $10^{13}$/sec. The activation energy for surface migration of isolated $Ga$ and $As$ atoms was chosen as 1.3 eV based on Ref.[7]. Based on Ref.[6], the activation energy for evaporation of an isolated $As$ was chosen as 1.675 eV.

The MBE growth parameters for this study were obtained from the experimental data given in Ref.[1]. The growth temperature was chosen in the range $723 - 873^o$K and the flux rate was set at 2 Å/sec. The cation to anion flux ratio employed for the study was in the range $1:10$ to $1:20$. It is noted that the (100) substrate surface employed in this study is flat without any steps.

The material and growth data discussed above were employed to calculate the model parameters according to the procedure detailed in section IVB of Ref.[17]. The model parameters were obtained as a function of growth temperature. The time evolution equations given by Eqs 8a and 8b with modifications discussed in Eq 4, and the boundary conditions corresponding a flat substrate described by Eq 14 of Ref.[17] were solved numerically on a CRAY YMP 2/216 at NSCEE, UNLV, Nevada. The CPU time for a typical growth of 20 Å of $GaAs$ was about 4 hours.

## 2.3 Results and Discussion

Concentration profiles were obtained as a function of time for various growth temperatures. They are shown for 823°K, 848°K and 873°K in Figure 2a-c, respectively, for flux ratio 1 : 10 and in Figure 3a-c for flux ratio 1 : 20. Below 823°K, the concentration profiles look similar to the profile of 823°K and therefore are not displayed. From Figures 2 and 3, it is observed that at and below 823°K, the growing surface is $As$-stabilized as is expected when the cation to anion flux ratio is 1 : 10 or higher. But, as the temperature increases, the growth surface becomes less $As$-stabilized. It is also observed that the time delay between the growth of the $Ga$ and subsequent $As$ layers is constant throughout the growth of the $As$ layer for temperatures below 823°K. Above 823°K, the time delay is larger at the start of growth of $As$ layer compared to the the completion of the layer. The time delay within the growth of a monolayer of $As$ decreases with time. This effect is prominent and larger for higher temperatures and lower flux ratios.

The above observations about the time evolution of the concentration profiles can be explained as follows. The growth of an $As$ layer is controlled by two surface processes; adsorption and evaporation. At temperatures lower than 823°K, the evaporation of surface $As$ is negligible and therefore, the growth rate is equal to the adsorption rate which is constant during the growth of a layer. Thus, the time delay during the growth of the layer is constant with time. If the temperature is above 823°K, the temperature is high enough that the evaporation of surface $As$ begins. The growth of an $As$ layer is now controlled by the competition between the adsorption and evaporation processes. The growth rate is the difference between the adsorption rate and evaporation rate. The adsorption rate is independent of the coverage, whereas, the evaporation rate critically depends on the coverage through the activation energy for evaporation which is the binding energy of the atom as discussed under Eq 5. The binding energy of a surface *As* pair increases with coverage. Therefore, at the start of the growth of an *As* layer, the binding energy of the *As* pair is the smallest possible and therefore, the evaporation rate is the largest as given by Eq 5. The growth rate given by the difference between the constant adsorption rate and the large evaporation rate is small. Therefore, there is a large time delay at the start of the growth. As the coverage increases, (i.e.), $n$ increases, the *As* atoms attain more nearest neighbors, and hence their binding energy increases which results in a decrease of the evaporation rate as given by Eq 5. Then, the growth rate increases with the coverage which results in a continuous decrease of the the time delay. The time delay is more at higher temperature, due to an increase in the evaporation rate and at lower flux ratio due to smaller adsorption rate. Thus, at lower temperatures and higher flux ratios, the surface appears more *As*-stabilized. The description of the surface processes is in complete agreement with the mechanisms proposed in Ref.[1] based on the experimental observations.

The intensity of a specular spot ($1^o$ off Bragg) of reflection high energy electron diffraction system with 10 kV electron beam was calculated using kinematical theory of electron diffraction as a function of growth time. The time averaged RHEED intensities, $TRI$(T) were calculated for various growth temperatures. A plot of $TRI$(T) versus growth temperature is shown in Figure 4 for flux ratios 1 : 10 and 1 : 20. The $TRI$(T) decreases below and above a certain called the transition temperature, and is identified as 770°K and 800°K, respectively, for flux ratios 1 : 10 and 1 : 20. The lower $TRI$(T) above and below the transition temperature is directly related to rougher surface. Below the transition temperature, the thermal activation for surface migration is low and therefore, the *Ga* and *As* atoms randomly adsorb on the surface at sites of their arrival resulting in a rough surface. As temperature increases towards the transition temperature, the thermal activation and hence surface migration increases, resulting in adatoms finding energetically more stable sites. This surface process decreases the surface roughness. Above the transition temperature, the evaporation of $As$ begins resulting in a rougher surface. The roughness of the surface directly correlates with decreased RHEED intensity due to destructive interference of the electron waves reflecting from various surface layers. Thus, the RHEED intensity peaks at the transition temperature. The flux ratio dependence of $TRI$(T) can be explained as follows. Lower flux ratio results in longer time for the formation of surface atom clusters with more than two $As$ atoms. This implies that the average evaporation rate during the growth of monolayer of $As$ is larger due lower coverage dependent activation energy for evaporation and hence lower transition temperature. This is in good agreement with the work of Chen et al. [1].

The experimental observations of steady state RHEED intensity were explained in Ref.[1] in terms of the competition between various surface processes such as adsorption and evaporation of atoms which are surface roughening processes and the atoms migrating to the step edges which is a surface smoothing process. This explanation is similar to the one given in this paper as discussed below. Surface roughening occurs at all temperatures due to random incorporation of atoms. At low temperatures, the surface migration is less and hence the surface smoothing effect is less. But, at high temperatures, the interlayer and intralayer surface migration aid in atoms reaching energetically favorable sites. Above the transition temperature, evaporation of $As$ from the surface is responsible for the rougher surface and lower RHEED intensity which is in agreement with the proposed mechanism in Ref.[1]. Thus, the stochastic model developed in Ref.[17-19] and used in this study for the MBE growth studies of $GaAs$ (100), accurately describes the kinetics of surface roughening of $GaAs$ and also aids in understanding the details of the mechanisms underlying the surface roughening phenomenon.

The plot of $TRI$ versus temperature obtained in this study was compared with that of the experimental study of Chen et al [1] and semi-quantitative agreement between the results was obtained. There are two main reasons for the quantitative differences between the results: (i) The molecular species is employed for the experiments ($As_4$) and our work ($As_2$) are different. (ii) the flux ratios employed in the experiment and our work may be different as experimental flux ratios are always reported in equivalent beam pressure ratios not in terms of rates of arrival of anion to cation as done in our work. Both of this can influence the result of transition temperature quantitatively.

### 2.4 Summary

The stochastic model of MBE growth based on the master equation approach with solid-on-solid restriction and quasi-chemical approximation employed for the study of surface processes in (100) $GaAs$ growth. The growth rate, the time averaged surface roughness and the time averaged RHEED intensity were obtained for various growth temperatures. The kinetic surface roughening transition temperature for the MBE growth of $GaAs$ is identified as 770°K and 800°K for flux ratios 1 : 10 and 1 : 20, respectively, from the temperature dependence of the time averaged RHEED intensity. The results of this study compare favorably with that of the experiments obtained under similar growth conditions[1]. The phenomenon of kinetic surface roughening transition in the MBE growth of GaAs (100) is explained in terms of the competition among various surface processes such as the incorporation and evaporation of atoms which roughen the surface and the surface migration of atoms to energetically favorable sites which smoothens the process.

## 3 Surface Ordering Kinetics of GaAlAs

### 3.1 Background

The presence of long range ordering in as-grown epilayers reduces the band gap of the material and thus has implications for opto-electronic device applications. Long range order has been observed in many compound semiconductors grown by MBE and MOCVD [21-22]. A few of the compound semiconductors which exhibit such order in as grown samples are: $GaAlAs$, $GaAsSb$, $InAsSb$, $GaInAs$ and $GaInP$. The presence of ordering in the epilayers is usually observed using transmission electron microscopy (TEM). The ordering observed in these compounds has been shown to be affected by growth and surface conditions and to be dependent on orientation. For example, in MBE grown $Ga_{1-x}A_xlAs$, it was observed that the degree of ordering as observed in terms of the intensity of superstructure reflections in TEM observations, depends on the substrate orientation, growth temperature and the $Al$ content [22].

In this work, the stochastic model approach developed for the study of the MBE growth of compound semiconductors [17-19] is employed to study the MBE surface ordering kinetics of $Ga_{1-x}Al_xAs$. In section 3.2, a brief discussion of the stochastic model for the MBE alloying studies is presented. The results of the surface ordering kinetic study of $Ga_{1-x}Al_xAs$ are discussed and compared with the experimental work of Ref.[21] in section 3.3. Summary is stated in section 3.4.

## 3.2 Stochastic Model for Alloy Kinetic Studies

The details of the development of the stochastic model for the MBE growth of alloy compound semiconductors has been presented elsewhere [17,18]. Due to limited space, only the salient features of the stochastic model are discussed in the following section.

### 3.2.1 Time Evolution Equations for MBE Kinetics

The stochastic model describes the time evolution of the macrovariables of growth in terms of the rates of the surface kinetic processes. The development of the model is based on the master equation scheme and the random distribution approximation. The assumptions necessary for the derivation are: (i) a rigid zinc blende lattice oriented along the 001 direction (ii) exclusion of the effects of surface reconstruction and strain (iii) exclusion of the creation of anti-site defects The kinetic process considered in the description of the time evolution equations are: adsorption and surface migration of $Ga$, $Al$ and $As$. All of the arriving cations ($Ga$ and $Al$) are allowed to adsorb at or the near the sites of their arrival depending on whether they arrive at a proper site. Since all the arriving cations absorb at the site of their arrival or a nearest neighbor site, the sticking coefficient of the cations is maintained at unity as experimentally observed. In the range of temperature of this study, the evaporation of atoms is assumed to be negligible. The rate of surface migration of atoms, R ($sec^{-1}$), is given by:

$$R = R_o \, e^{(E_{act}+xK_{aa})/kT} \tag{6}$$

where $R_o$ is the frequency factor and the $E_{act}$ the activation energy for an isolated atom on the terrace, $x$ the number of in-plane atom-atom bonds and $K_{aa}$ the interaction energy of such a bond. Both the intralayer and interlayer surface migrations rates are assumed equal.

### 3.2.2 Macrovariables

Two sets of macrovariables - one for each sublattice can be defined. For the purpose of this study, it is assumed that the cations belong to the even sublattice and the anion, $As$, belongs to the odd sublattice. The macrovariables for the $2n^{th}$ layer are: concentration variables, $C_{Ga}$(2n) and $C_{Al}$(2n), atom-vacancy bond densities, $Q_{Ga}$(2n) and $Q_{Al}$(2n) and atom-atom bond densities, $\tilde{N}_{GaAl}$(2n), $\tilde{N}_{GaGa}$(2n) and $\tilde{N}_{AlAl}$(2n). All the bonds referred to in this manuscript are second nearest neighbor bonds, where the whole crystal is considered. They are also the first nearest neighbor inplane bonds in the (001) plane. Of the seven macrovariables, only five are independent because of the following relations:

$$\tilde{N}_{GaGa}(2n) = 2\, C_{Ga}(2n) - \frac{1}{2} Q_{Ga}(2n) - \frac{1}{2} \tilde{N}_{GaAl}(2n)$$

for $Ga - Ga$ bond density,

$$\tilde{N}_{AlAl}(2n) = 2\, C_{Al}(2n) - \frac{1}{2} Q_{Al}(2n) - \frac{1}{2} \tilde{N}_{GaAl}(2n)$$

for $Al - Al$ bond density and

$$\tilde{N}_{vv}(2n) = 2\,(C_v(2n)) - \frac{1}{2}Q(2n) \tag{7}$$

where

$$C_v(2n) = (1.0 - C(2n))$$

and

$$C(2n) = C_{Ga}(2n) + C_{Al}(2n)$$

and

$$Q(2n) = Q_{Ga}(2n) + Q_{Al}(2n)$$

where $C_v$(2n) is the vacancy density, and $C(2n)$ is the total atom concentration in the $2n^{th}$ layer. In Eq. 2, it is assumed that the inplane coordination number is four. Similar set of equations can be written for the anion sublattice. For the study $Ga_{0.5}Al_{0.5}As$, only the cation sublattice is alloyed and therefore, there are two kinds of atoms- $Ga$ and $Al$ are present in the cation sublattice and only $As$ atoms present in the anion sublattice. There will be five independent variables for the cation sublattice and two independent variables for the anion sublattice. Thus, there is a total of seven independent macrovariables whose time evolution needs to be modeled for a complete description of the MBE growth kinetics. In this study, the independent variables for the cation sublattice are chosen as: $C_{Ga}$(2n) $C_{Al}$(2n), $Q_{Ga}$(2n), $Q_{Al}$(2n) and $\tilde{N}_{GaAl}$(2n). For the anion sublattice, the independent variables are: $C_{As}$(2n+1) and $Q_{As}$(2n+1).

### 3.2.3 The $Ga_{1-x}Al_xAs$ Alloy System and Model Parameters

The $Ga - Ga$, $Al - Al$ and $Ga - Al$ second nearest neighbor pair interaction energies were obtained from the first principle calculations and found to be [6]:

$$V_{Ga-Ga} = 0.000eV,$$

$$V_{Al-Al} = 0.000eV$$

$$V_{Ga-Al} = 0.134eV \tag{8}$$

It is noted that the $V_{Ga-Ga}$, $V_{Al-Al}$, and $V_{Ga-Al}$ are different for *sigma* and *pi* bonds (i.e.), if they are along the covalent bond or perpendicular to it. Since, the present model is unable to differentiate between the *sigma* and *pi* bonds in a layer, the energy values are averaged and used for both types of bonds. This assumption is one of the possible reasons that the present study may show the correct kinetics, but may not show the correct type of ordering in the grown crystal. The activation energy of surface migration for *Ga* and *Al* are assumed to be 1.3 eV and the frequency factor is assumed to be 1.0 x $10^{13}$/sec. The growth conditions considered for this study are: flux rate of cations is 2 Å/sec; cation to anion flux ratios 1 : 10, 1 : 20 and 1 : 30; and substrate temperatures in the range of 760 – 880°K.

Since the equations governing the time evolution of the macrovariables are coupled, nonlinear, first-order differential equations, they are not analytically integrable. They were numerically integrated using a numerical integration scheme on a CRAY YMP 2/216. The average computational time for a typical growth of 7 monolayers was 3-4 CPU hours.

## 3.3 Results and Discussion

The macrovariables, $C_{Ga}$(2n), $C_{Al}$(2n), $C_{As}$(2n+1), $Q_{Ga}$(2n), $Q_{Al}$(2n), $Q_{As}$(2n+1), and $\tilde{N}_{GaAl}$(n) were obtained as a function of time for various temperatures in the range of 760-880°K which is the typical range of growth for these compounds. The short range order (SRO) parameter defined by:

$$SRO = \sum_{i=1}^{i=3} \left( \frac{\tilde{N}_{GaGa}(2n) + \tilde{N}_{AlAl}(2n) - \tilde{N}_{GaAl}(n)}{C(2n)} \right) \quad (9)$$

was evaluated or various flux ratios, temperatures and arsenic species. The SRO defined in Eq. 4 is 0.0, 1.0 and 2.0 for completely seggregated, completely random and completely ordered alloys, respectively.

A plot of SRO parameter versus temperature is shown in Figure 1 for the case of the diatomic molecular species $As_2$ (cracked arsenic) with flux ratios of 1 : 10, 1 : 20 and 1 : 30. The kinetic order-disorder temperature seen from Figure 1 to be 770°K, 800°K and 810°K for flux ratios of 1 : 10, 1 : 20 and 1 : 30, respectively. The temperature dependence of the SRO parameter can be explained as follows. In the low temperature regime, the effective migration rate of *Ga* and *Al* atoms is small and therefore, their chances of sampling many different surface atomic configurations is limited. Thus, these atoms incorporate randomly at their arrival sites leading to a smaller SRO parameter. As the temperature increases, the effective migration rate increases resulting in more sampling of various surface configurations by the *Ga* and *Al* atoms which enables them to find energetically favorable atomic sites more efficiently. At and above the transition temperature, the thermal energy becomes comparable to the energy difference between the energetically most and least favorable surface atomic configuration. Thus, the thermal randomization of surface atomic configurations sets in and leads to a more random arrangement of *Ga* and *Al* atoms, resulting in a decrease of the SRO parameter with temperature. The transition temperature and the maximum value of the SRO parameter are lower for lower flux ratios because the effective migration rate for lower flux ratios are higher.

A plot of the temperature dependence of the Isolated Terrace Cation Parameter (ITCP) (ITCP is the concentration of isolated terrace cation which can be obtained from the concentrations of cations, atom-vacancy bond density and atom-atom bond density using randon distribution approximation as defined in

Ref. 19) is shown in Figure 2. The ITCP increases with temperature for flux ratios 1 : 10, 1 : 20 and 1 : 30. This behavior indicates that the transition temperature is close to 760°K which is consistent with the results of Figure 1. The physical explanation of thermal randomization increasing the concentration of isolated terrace adatoms correlates well with decreasing degree of ordering.

Similar trends in the temperature dependence were observed for the case of monatomic *As*. It is not reported here due to limited space. These results were compared with the experimental work reported in Ref.[21]. The qualitative agreement between the results in terms of the temperature dependence of the SRO parameter is excellent. One of the major differences between the results is that the type of ordering reported in Ref.[21] for the MBE growth of (100) $Ga_{0.5}Al_{0.5}As$ Experimentally, $L1_o$ type ordering is observed. Our results show the ordered structure in which *Ga* atoms are surrounded by *Al* atoms and vice versa. The above described structure is not an $L1_o$ type ordered structure. The reasons for this discrepancy in the type of ordering are as follows. Firstly, the $L1_o$ type ordering results from the difference between the bulk-surface and surface-surface pair interaction energies of atoms which the stochastic model does not take into account. Secondly, the atom pair interaction energies for *sigma* and *pi* bonds were averaged out and both bonds were treated equal for simplicity. At present, the authors are working on incorporating this detail into the stochastic model.

### 3.3.1 Summary

The surface kinetics of MBE growth of (100) $Ga_{0.5}Al_{0.5}As$ was studied theoretically using the stochastic model for various growth conditions. The degree of ordering was obtained in terms of the short range order (SRO) parameter. The order-disorder temperature was obtained for flux ratios of 1 : 10, 1 : 20 and 1 : 30. Both *As* and $As_2$ species were considered. For the $As_2$ soource, the order-disorder temperature was found to be 780°K with a maximum degree of ordering of 85 % for a flux ratio of 1 : 10; 800°K with a maximum degree of ordering of 84 % for a flux ratio of 1 : 20; and 810°K with a maximum degree of ordering of 83 % for a flux ratio of 1 : 30. The results are in good qualitative agreement with experiments. The surface ordering kinetics observed can be described in terms of the effective surface migration rate of cations as follows. The qualitative agreement between experiments and theory is excellent. The key difference between the experiments and theory is in the type of ordering observed. This difference is attributed to the indistinguishability of various types of second nearest neighbor bonds, (i.e.), *sigma*, *pi*, surface-surface and surface-bulk, in the present stochastic model. Presently, work is in progress to incorporate these differences in the bonds in to the stochastic model.

# 4 Doping Kinetics in Semiconductors

## 4.1 Background

Doping an epilayer during its growth by coevaporating the dopant has been the standard practice in molecular beam epitaxy (MBE) of semiconductors. Due to low growth rates (1 $\mu$m/hr.) and low temperature of growth (i.e., limited or negligible bulk diffusion), abrupt dopant profiles should be achievable in MBE grown epilayers. However, for many dopants in $Si$ MBE growth, either a smearing of the dopant profile and/or surface enrichment of the dopant have been observed experimentally [24]. In addition, it has been observed that some dopants incorporate inefficiently in some semiconductors due to excessive evaporation dictated by low binding energy of these dopants to the growing surface.

Many theoretical models, both thermodynamic and kinetics based, have been proposed to explain these experimental observations. Iyer et al [25] proposed a kinetic model which described the time evolution of surface concentration of dopant in terms of incorporation and evaporation. Their model was good enough to explain many experimental observations in the doping of $Si$ and $GaAs$ [25]. However, it did not explicitly include the surface segregation phenomenon. Barnett et al [26] proposed a model including the segregation phenomenon in which they assumed that dopants from several nanometers of subsurface diffuse to the surface. Rockett et al [27] proposed a kinetic model to describe the surface segregation phenomenon of $Sn$ in $GaAs$. In their approach, a relationship between the bulk diffusion coefficient and surface diffusion coefficient was assumed. This relationship does not include effects of surface conditions such as the roughness or coverage. From the analysis and comparison of the results of their model with the experiments, they concluded that their model was sufficient to explain the experimental observations. Andrieu et al [24] proposed new physical mechanism of dopants climbing to the surface and developed a rate equation model based on this assumption. Their model was able to explain the surface segregation phenomenon satisfactorily. However, the pre-exponential factor for their rate processes was $10^{-11}s^{-1}$ which is not physically justifiable.

Even though each of these model is suitable for one or more dopant-semiconductor systems, none of them is, however, completely satisfactory for all systems, incorporate all observed effects. The model developed by Andrieu et al [24] appears to be the best available model, it employs physically unjustifiable exponential prefactors and therefore is questionable. This manuscript proposes a general rate equation model which captures most aspects of the in-situ doping kinetics and employs physically reasonable and justifiable parameters. Thus, it overcomes most of the limitations of the earlier proposed models.

The phenomenological rate equation model is presented in section 4.2. A discussion of the model parameters is presented in section 4.3. The results of the application of this model to $In$ doping of $Si$ are presented and compared with available experimental data in section 4.4. A discussion of the results is also presented in section 4.4. Summary is presented in section 4.5.

## 4.2 Rate Equation Model for Doping

The elementary surface kinetic processes that control the doping kinetics are: adsorption, evaporation and interlayer migration of the the host atom, $Si$, and the dopant, $In$, ($In$ is chosen for this study, but the model is general and can be applied to any system). The rate of change of concentration of $Si$ in the $n^{th}$ layer can be written in terms of the rates of microscopic kinetic processes such as discussed above. Describing the time evolution of the concentration of $Si$ in the $n^{th}$ layer, $C(n)$, we get:

$$\begin{aligned}
\frac{dC_{Si}(n)}{dt} &= J_{Si}\,[C(n-1)-C(n)] && (A)\\
&- R_o e^{\frac{-E_{ave,eva,Si}}{kT}}\,[C(n)-C(n+1)]\left(\frac{C_{Si}(n)}{C(n)}\right) && (B)\\
&+ R_o e^{\frac{-E_{ave,dif,Si}}{kT}}\,[C(n-1)-C(n)]\\
&\text{x}\ \left(\frac{C_{Si}(n+1)}{C(n+1)}\,[C(n+1)-C(n+2)]\,\frac{C_{Si}(n-1)}{C(n-1)}\,[C(n-1)-C(n)]\right) && (C)\\
&- R_o e^{\frac{-E_{ave,dif,Si}}{kT}}\,[C(n)-C(n+1)]\left(\frac{C_{Si}(n)}{C(n)}\right)\\
&\text{x}\ \left([C(n+1)-C(n+2)]\,[C(n-1)-C(n)]\right) && (D)
\end{aligned} \tag{10}$$

Term $A$ describes the rate of adsorption of $Si$ on to the $n^{th}$ layer in terms of the available sites for adsorption, $[C(n-1)-C(n)]$, and the flux rate, $J_{Si}$. The description in Term $A$ assumes that the atoms adsorb with unity sticking probability at sites which are available for adsorption. Term $B$ describes the rate of loss of $Si$ from the $n^{th}$ layer due to evaporation in terms of the number of atoms exposed to vapor, $\frac{C_{Si}(n)}{C(n)}\,[C(n)-C(n+1)]$, the frequency factor which is assumed to be $10^{13}$/sec. and the average activation energy for evaporation of $Si$ atoms which depends on the concentration of the layer. In this description, it is assumed that number of nearest neighbors for a $Si$ atom increases directly as the concentration of $Si$ in the layer. Note that since the dopants are in $ppm$ levels that they do not affect the binding energy of $Si$ atoms. Thus, the average activation energy for evaporation of a $Si$ atom in the $n^{th}$ layer, $E_{ave,eva,Si}$, is given by:

$$E_{ave,eva,Si} \;=\; E_{iso,eva,Si} \;+\; zE_{SiSi}C_{Si}(n) \tag{11}$$

where $z$ is the inplane coordination number which is 4 for (001) plane and 6 for (111) plane and $E_{SiSi}$ and $E_{iso,eva,Si}$ are the second nearest neighbor interaction energy and the activation energy for the evaporation of an isolated $Si$, respectively. (Note that in the (001) and (111) planes, the nearest neighbor atoms are actually the second nearest neighbor atoms when the whole crystal is considered.) Term $C$ in Equation 1 is the rate of gain of $Si$ atoms in the $n^{th}$ layer due to interlayer migration of $Si$ to the $n^{th}$ layer from the adjacent layers, $n-1$ and $n+1$. Term $D$ in Equation 1 is the rate of loss of $Si$ atoms in the $n^{th}$ layer due to interlayer migration of $Si$ from the $n^{th}$ layer to the adjacent layers, $n-1$ and $n+1$. The average activation energy for the migration of $Si$, $E_{ave,dif,Si}$ is given by:

$$E_{ave,dif,Si} = E_{iso,dif,Si} + zE_{SiSi}C_{Si}(n) \tag{12}$$

where $E_{iso,dif,Si}$ is the activation energy for interlayer migration of an isolated $Si$ atom.

Similar equation for the rate of change of dopant concentration in the $n^{th}$ layer can be written by simply replacing $Si$ with $In$ where $In$ stands for the dopant. In the rate equation for the dopant, the activation energy for evaporation and migration should be redefined as:

$$E_{ave,eva,In} = E_{iso,eva,In} + zE_{SiIn}C_{Si}(n) \tag{13}$$

and

$$E_{ave,dif,In} = E_{iso,dif,In} + zE_{SiIn}C_{Si}(n) \tag{14}$$

respectively. In the description of Equations 4 and 5, the fact that the second nearest neighbors of dopant atoms are essentially $Si$ atoms due to the $ppm$ level concentrations of dopant utilized.

## 4.3 Model Parameters and Growth Parameters

The activation energy for the evaporation of an isolated $Si$, $E_{iso,eva,Si}$, can be obtained for (100) and (111) growths as 2.6 eV and 3.2 eV, respectively [20]. The second nearest neighbor interaction energy, $E_{SiSi}$ for (100) and (111) growths are estimated to be 0.25 eV and 0.325 eV, respectively from Ref.[20]. The activation energy for evaporation of $In$ is not reported in the literature. Thus, a trial and error approach was taken to fit two of the data points on the concentration of $In$ versus $\frac{1}{T}$ plot reported in Ref.[28]. Thus, we obtained a value of 1.6 eV and -0.25 eV for $E_{iso,eva,In}$ and $E_{SiIn}$, respectively. The activation energy for migration of an isolated $Si$, $E_{iso,dif,Si}$, was chosen as 0.8 eV based on available experimental data for activation energy for $Si$ [7]. $E_{iso,dif,In}$, was assumed to be equal to that of $Si$ for lack of available data.

The growth parameters were chosen exactly as used for the experimental work reported in Ref.[24] and are presented below. The flux rate of $Si$, $J_{Si}$, was kept at 1$\mu$m/hr. The flux ratio between $Si$ and $In$, $J_{Si}/J_{In}$, was maintained at 2 x $10^{-4}$. The evaporation of $Si$ was negligible in the temperature range of this study and therefore the growth rate of the epilayer was approximately equal to the $J_{Si}$. The substrate temperature was in the range 500 – 750°C.

The differential equations given by Equation 1 and a similar one for $In$ are coupled non-linear first order differential equations which are not analytically integrable. Therefore, these equations were solved numerically on the CRAY YMP2/245 located at National Supercomputer center for Energy and Environment (NSCEE) at University of Nevada, Las Vegas. The boundary condition is that the first 3 layers are full with $Si$ and rest of the layers are empty at the start of the growth which corresponds to assuming a flat substrate. A typical run for a growth of 20 seconds took about 1 CRAY CPU hour.

## 4.4 Results and Discussion

Plots of dopant sticking coefficient, $S_{In}$, versus time (sec.) for various growth temperatures of the study were obtained and analyzed. (The growths were simulated only for about 20 seconds due to computer time limitations.) Analysis of the data indicated that $S_{In}$ is an exponentially decaying function of time. Exponentials of the form, $A(T)e^{-\frac{t}{\tau(T)}}$, were fitted for each temperature. Note that $A(T)$ and $\tau(T)$ are functions of temperature. Using $A(T)$ and $\tau(T)$, $S_{In}$ was obtained for the case of 3000 seconds of growth by extrapolation. Such an extrapolation is justified because the growth rate is constant and the layer-by-layer growth mode is maintained for all growth conditions of in this study.

A plot of the extrapolated $S_{In}$ versus $\frac{1}{T}$ is shown in Figure 1 along with the experimental data. The agreement between the theoretical values and the experimental values from Ref.[28] is excellent for the entire temperature range. The mechanism which results in the temperature dependence of $S_{In}$ is as follows. At low temperatures, the surface segregation aided evaporation of $In$ (due to its repulsive interaction with the host sublattice) is not dominant as the thermal energy is not enough to allow this activated process. As temperature increases, the interlayer migration rate of $In$ to the surface increases and the $In$ concentration increases in the surface layer. More $In$ in the top surface layer, results in more opportunities for evaporation. Thus, a larger evaporation of $In$ results at higher temperature. The evaporation rate of $In$ aided by the surface segregation process is much larger than the typical evaporation rate of atoms from the surface.

Plots of the dopant concentration, $C_{In}$, versus distance from the surface are shown in Figure 2 for various growth times for a growth at 660°C. It is observed that the $In$ concentration in the bulk is about the same and is independent of the time of growth. The surface concentration of $In$, however, is increasing with time as expected. Note that there is a dip in $C_{In}$ for all the profiles just below the surface layer. The concentration of $In$ in this zone is an order of magnitude less than that in the bulk and many orders of magnitude less than that at the surface. We call this region the dopant depleted zone (DDZ). The physical reason for this phenomenon is as follows. The dopant, $In$, segregates from the layers below upwards due to the repulsive atomic interaction with the host lattice. In other words, considering the $n^{th}$ layer, $In$ atoms migrate from the $n-1^{th}$ layer to the $n^{th}$ layer which increases $C_{In}(n)$ and to the $n+1^{th}$ layer from the $n^{th}$ layer which decreases $C_{In}(n)$. The rates of these processes depend on the availability of $In$ atoms in the respective layers which are exposed to vapor so that they can migrate and the availability of sites in the respective layers. For the migrations to the $n^{th}$ layer compared to the migrations from the $n^{th}$, both of these factors are small. Thus, the rate of migration to the surface layer from the subsurface layer is larger than rate of migration of $In$ to the subsurface layer from one layer below. This difference in the rates results in a deficiency of $In$ atoms in the subsurface layer. This resembles the phenomenon of a precipitation depleted zone (PDZ) which result near the grain boundaries in many alloys. This result was compared with the experimental data and a careful observation of the segregation profile data shown in Figure 2 of Ref.[28] does show that there is a dip in the profile. In our result the dip is much more pronounced compared to that of the experiments which may be due to stronger repulsive interaction energies used in our model.

Plots of the $In$ segregation profiles for 853°K, 893°K, and 933°K are shown in Figure 3. It appears that the segregation profiles are similar for various temperatures except that the bulk $In$ concentrations are lower for higher temperature which correlates well with Figure 1. This type of a dopant depleted zone may not always be present even if the dopant interactions with the host lattice is repulsive. It depends on a variety of factors such as the growth rate, the strength of the repulsive interaction ($E_{SiIn}$), the ratio, $J_{Si}/J_{In}$ and others.

## 4.5 Summary

A rate equation model based on the master equation is developed for the study of MBE doping kinetics. The model is applied to study the surface segregation phenomenon during $In$ doping of $Si$. The doping studies were performed for various growth conditions. The results of the predicted sticking coefficient of $In$

versus $\frac{1}{T}$ and the dopant depth profile obtained shows excellent agreement with experiments. The sticking coefficient decreases with $T$ due to surface segregation aided evaporation of $In$ at higher temperatures. The surface segregation of $In$ occurs due to strong repulsive interaction between $In$ and the host lattice which results in upward migration of $In$. A dopant depleted zone where the $In$ concentration is lower than that in the bulk and at the surface is observed and agrees well with experiments.

## 5 Conclusion

The stochastic modeling of MBE growth was employed to study three growth kinetics problems: surface roughening kinetics of $GaAs$, surface ordering of $GaAlAs$ and doping kinetics of semiconductors. The results obtained were compared with that reported in the literature and the agreement obtainmed was good. This study validates the use of the stochastic modeling as a viable and effective tool for MBE growth kinetic studies. Many spin off projects from this project are underway at present.

## Acknowledgement

The authors would like to acknowledge the partial financial support provided by AFOSR under contract # F 49620-90-C-09076 and CRAY Research Co. The author like to gratefully acknowledge the contributions of Dr. Srinivasan Krishnamurthy to this work.

### REFERENCES

1. P.Chen, J.Y.Kim and A.Madhukar, ''*Optimal surface and growth front of III-V semiconductors in molecular beam epitaxy: A study of kinetic processes via reflection high energy electron diffraction specular beam intensity measurements on GaAs (100)*'' J. Vac. Sci. Tech., vol B 4, (1986), p890.

2. C.T.Foxon, M.R.Boudry and B.A.Joyce, Surface Sci., vol 44, (1974), p69.

3. C.T.Foxon, and B.A.Joyce, Surface Sci., vol. 50, (1975) p434.

4. C.T.Foxon, and B.A.Joyce, Surface Sci., vol. 64, (1977), p293.

5. J.R.Arthur, J. Appl. Phys., vol. 39, (1968), p4032.

6. J.R.Arthur, Surface Sci., vol. 43, (1974), p449.

7. J.H.Neave, P.J.Dobson and B.A.Joyce, Appl. Phys. Lett., vol. 47, (1985), p100.

8. Seiichi Nagata and Tsuneo Tanaka, J. Appl. Phys., vol. 48, (1977), p940.

9. A.Madhukar and S.V.Ghaisas, CRC Crit. Rev. Solids State Mater. Sci., vol. 14, (1988), p1. (and references therein).

10. J.Singh and A.Madhukar, J. Vac. Sci. Tech., vol. 1, (1983), p385.

11. A.Madhukar and S.V.Ghaisas, Appl. Phys. Lett., vol. 47, (1983), p247.

12. J.Singh and K.K.Bajaj, J. Vac. Sci. Tech., vol. 3, (1986), p520.

13. J.Singh and K.K.Bajaj, Appl. Phy. lett., vol. 47, (1985), p594.

14. S.Clarke and D.D.Vvedensky, Phys. Rev. Lett., vol. 58, (1987), p2235.

15. D.D.Vvedensky and S.Clarke, Surface Sci., vol. 225, (1990), p373. (and references therein).

16. E.Chason, J.Y.Tsao, K.M.Horn and S.T.Picraux, J. Vac. Sci. Tech., B7(2), 332, (1989).

17. R.Venkatasubramanian, J. Matl. Research., 7, 1222, (1992).

18. R.Venkatasubramanian, J. Matl. Research., 7, 1236, (1992).

19. R.Venkatasubramanian and D.L.Dorsey, J. Vac. Sci. Tech., (to appear in Mar/Apr issue 1993.).

20. Srinivasan Krishnamurthy, M.A.Berding, A.Sher and A.-B.Chen, J. Appl. Phys., vol. 68, (1990), p4020.

21. T.S.Kuan, T.F.Kuech, W.I.Wang and W.L.Wilkie, Phys. Rev. Letts., 54, 1985, p201.

22. G.B.Stringfellow, J. Vac. Sci. Tech., **B9**, 1991, p2182 (and all the references therein).

23. S.Krishnamurthy, (private communication). (These values were obtained from first principle quantum mechanical calculations. Related article is in preparation for Phys. Rev. Lett.).

24. A. Andrieu, F. Arnaud d'Avitaya and J.C. Pfister, J. Appl. Phys., **65**, 2681, (1989).

25. S.S. Iyer, R.A. Metzger, and F.G. Allen, J. Appl. Phys., **52**, 5608, (1981)

26. S.A. Barnett and J.E. Greene, Surf. Sci., **151**, 67, (1985).

27. A. Rockett, S.A. Barnett, J.E. Greene, J. Knall, and J.E. Sundgren, J. Vac. Sci. Technol., **A 3**, 855, (1985).

28. J. Knall, J.E. Sundgren, J.E. Greene, A. Rockett and S.A. Barnett., Appl. Phys. Lett., **45**, 689, (1984).

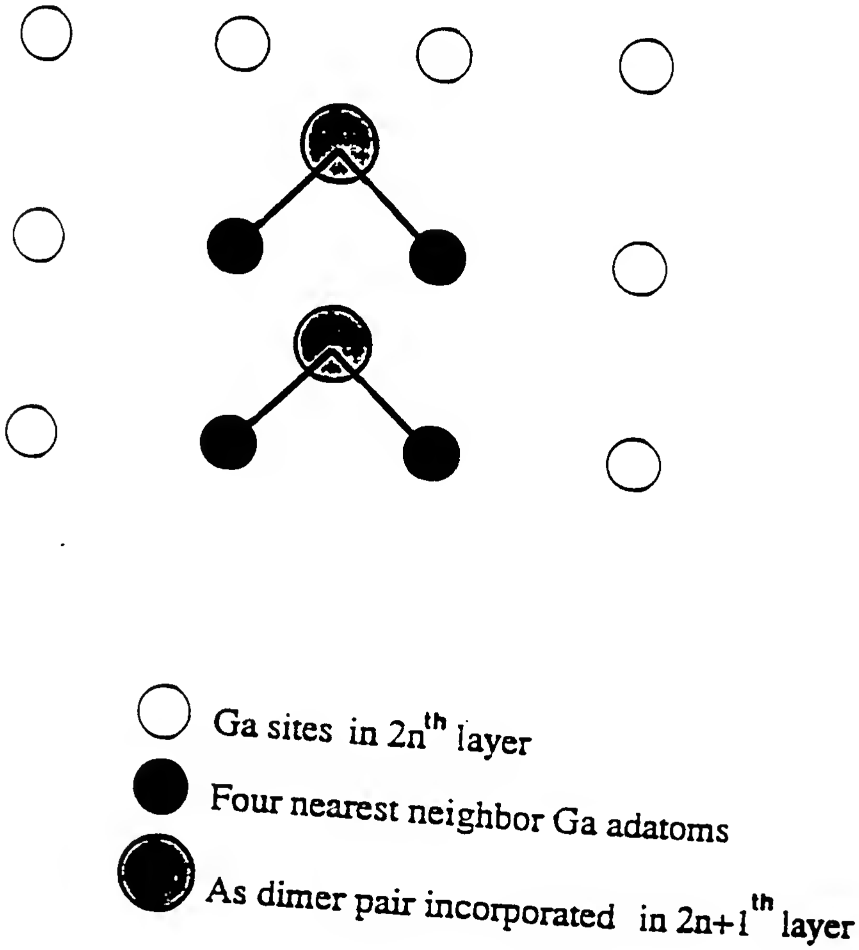

Ga sites in 2nth layer
Four nearest neighbor Ga adatoms
As dimer pair incorporated in 2n+1th layer

Fig. 1. Surface *Ga* atomic arrangement required for the incorporation of diatomic $As_2$.

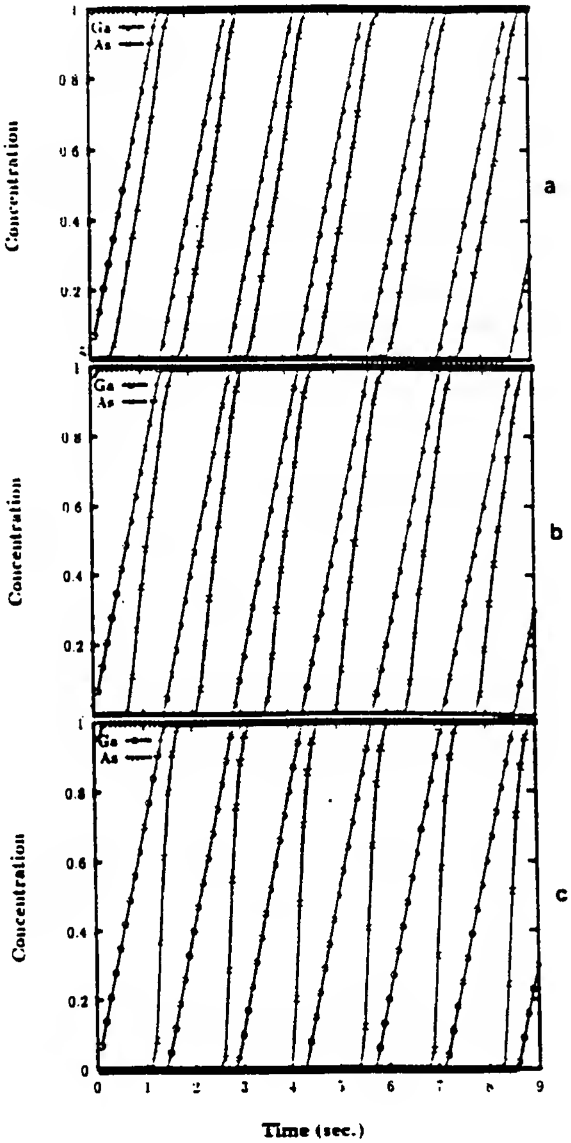


Fig. 2. Concentration profiles of *Ga* and *As* for flux ratio 1 : 10 for various temperatures. (a) 823°K (b) 848°K (c) 873°K.

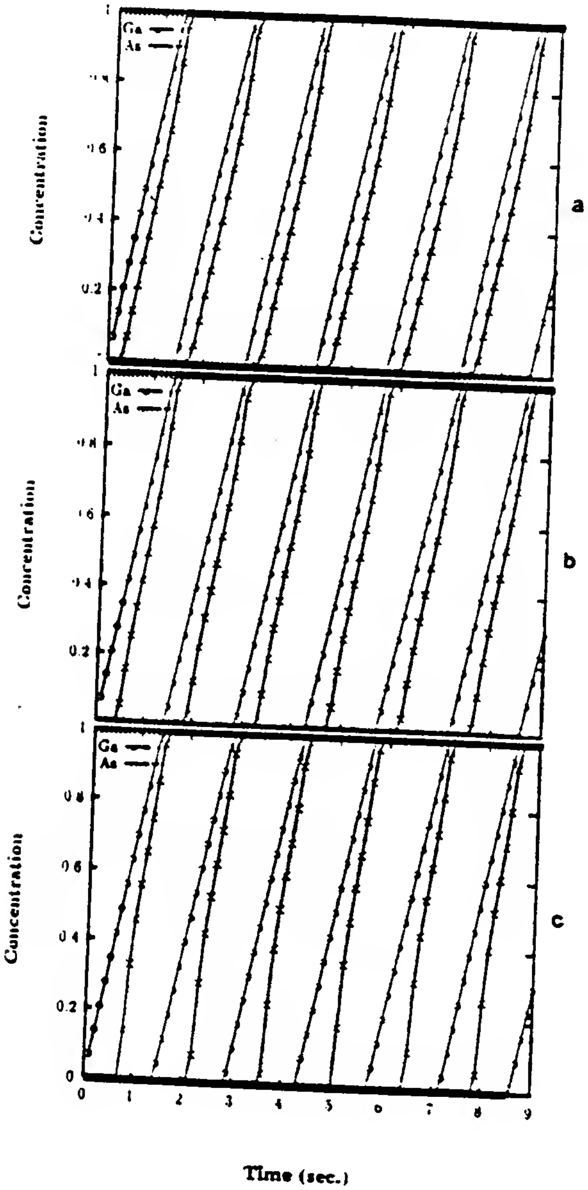


Fig. 3. Concentration profiles of $Ga$ and $As$ for flux ratio 1 : 20 for various temperatures. (a) 823°K (b) 848°K (c) 873°K.

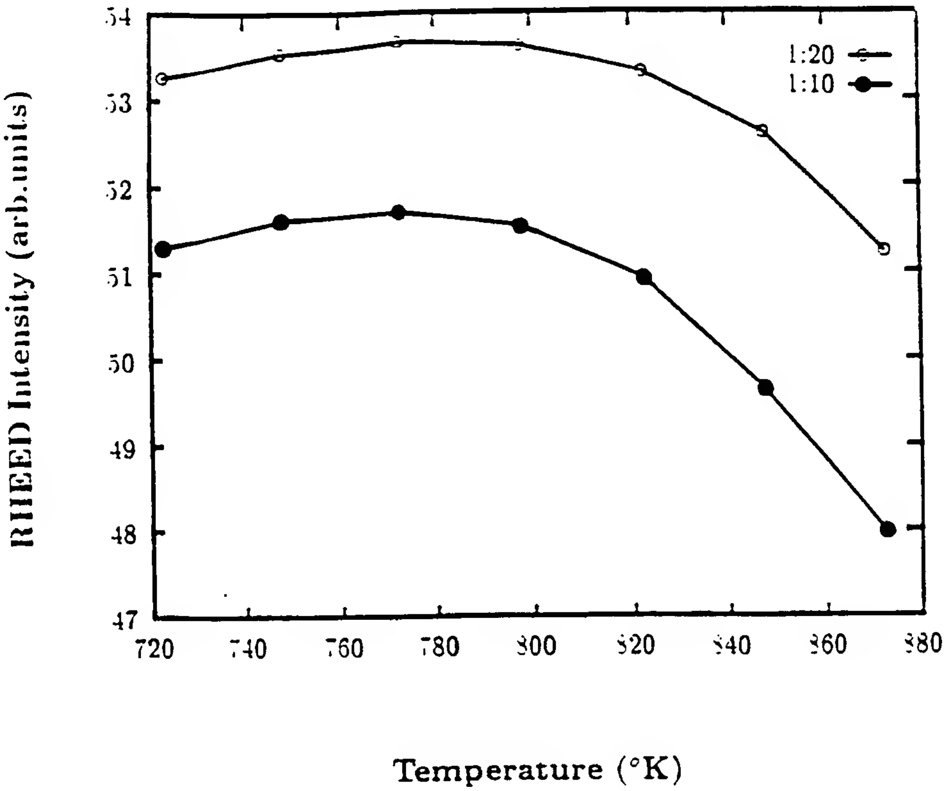


Fig. 4. Time averaged RHEED intensity, $TRI$(T) versus temperature for flux ratios 1 : 10 and 1 : 20.

Fig. 4. Time averaged RHEED intensity, $TRI$(T) versus temperature for flux ratios 1 : 10 and 1 : 20.

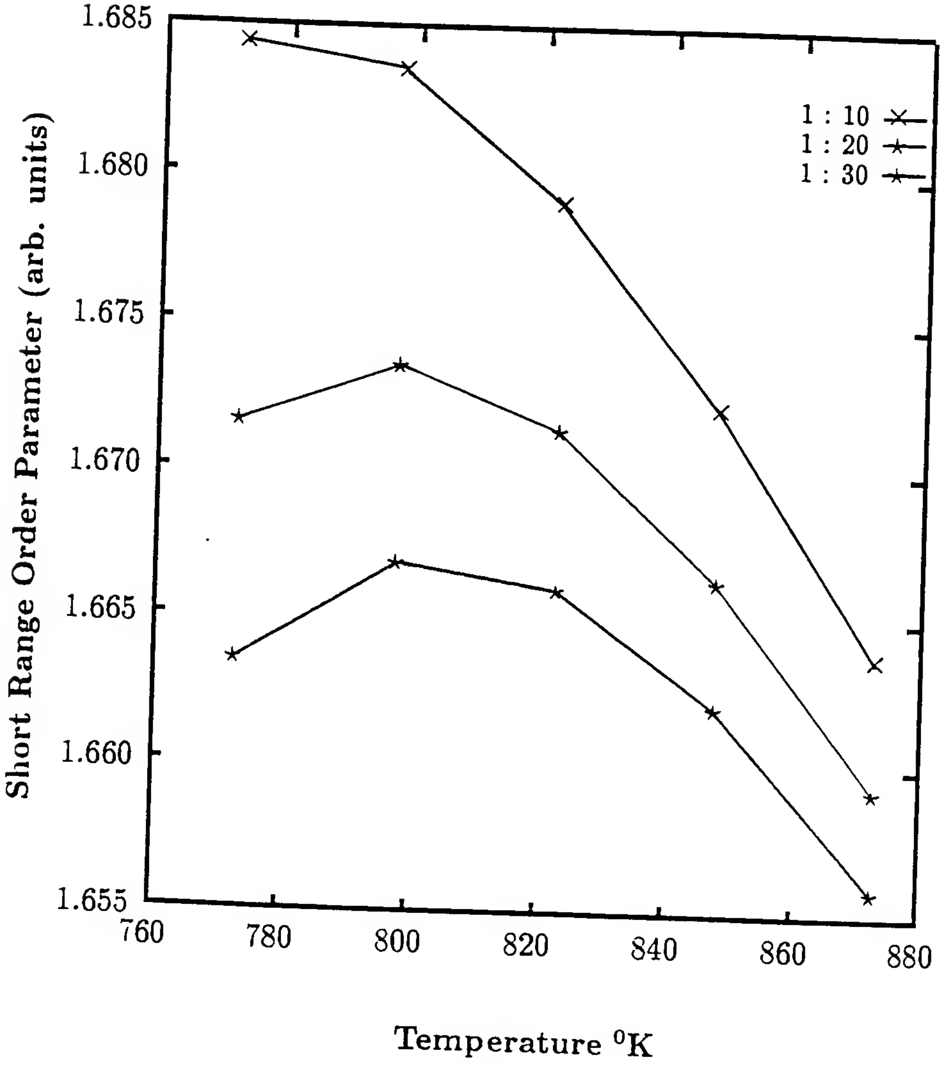


Fig. 5. A plot of SRO parameter versus temperature for flux ratios 1 : 10, 1 : 20 and 1 : 30.

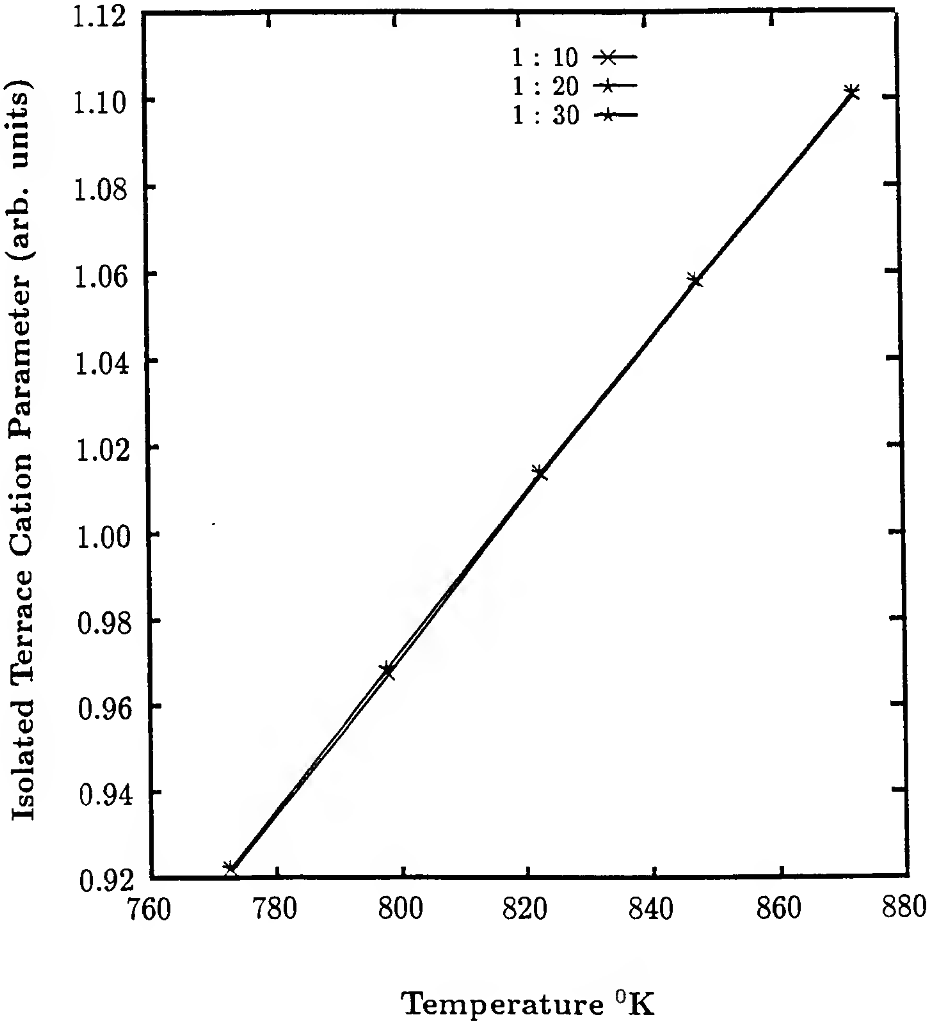


Fig. 6. A plot of isolated terrace adatom parameter versus temperature for flux ratios 1 : 10, 1 : 20 and 1 : 30.

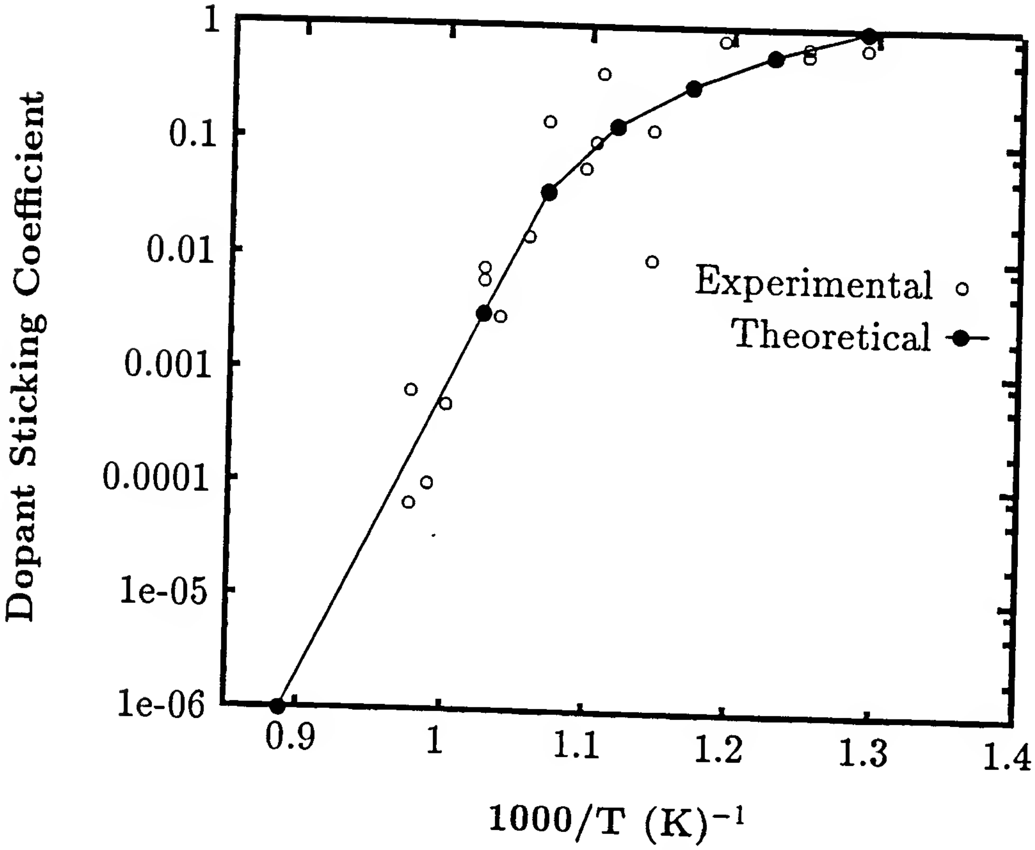


Fig. 7. Plot of sticking coefficient of dopant, *In*, versus $\frac{1}{T}$.

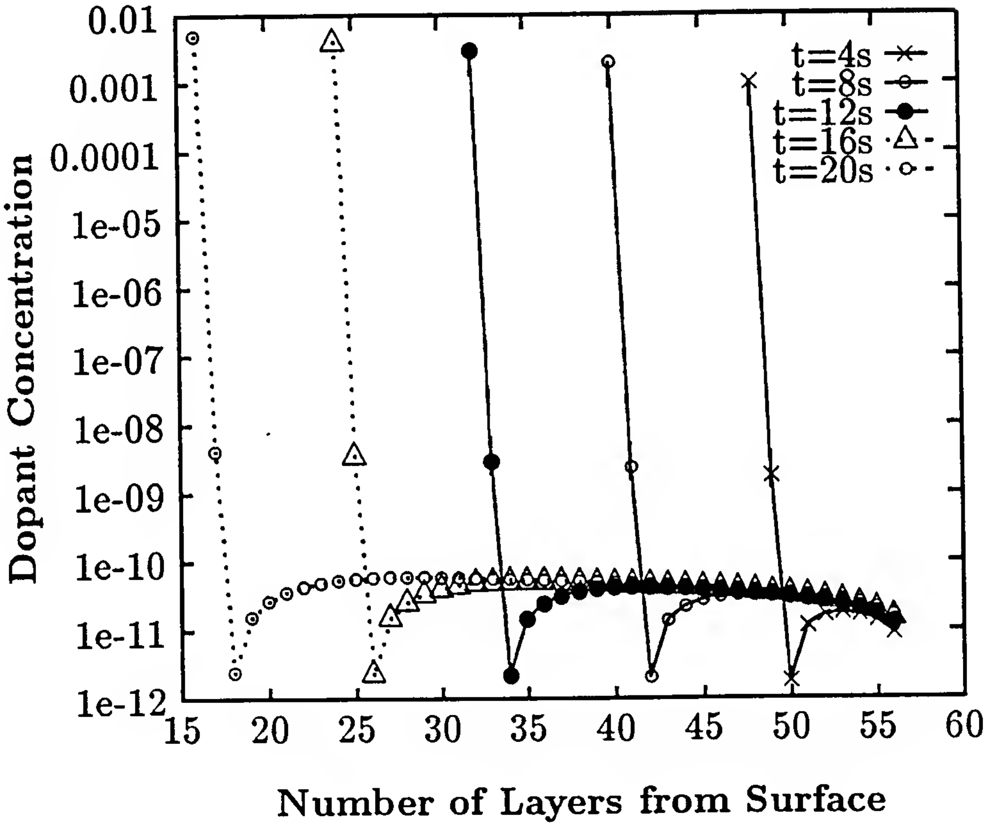


Fig. 8. Plots of dopant segregation profile for various growth times for the growth at 933°K.

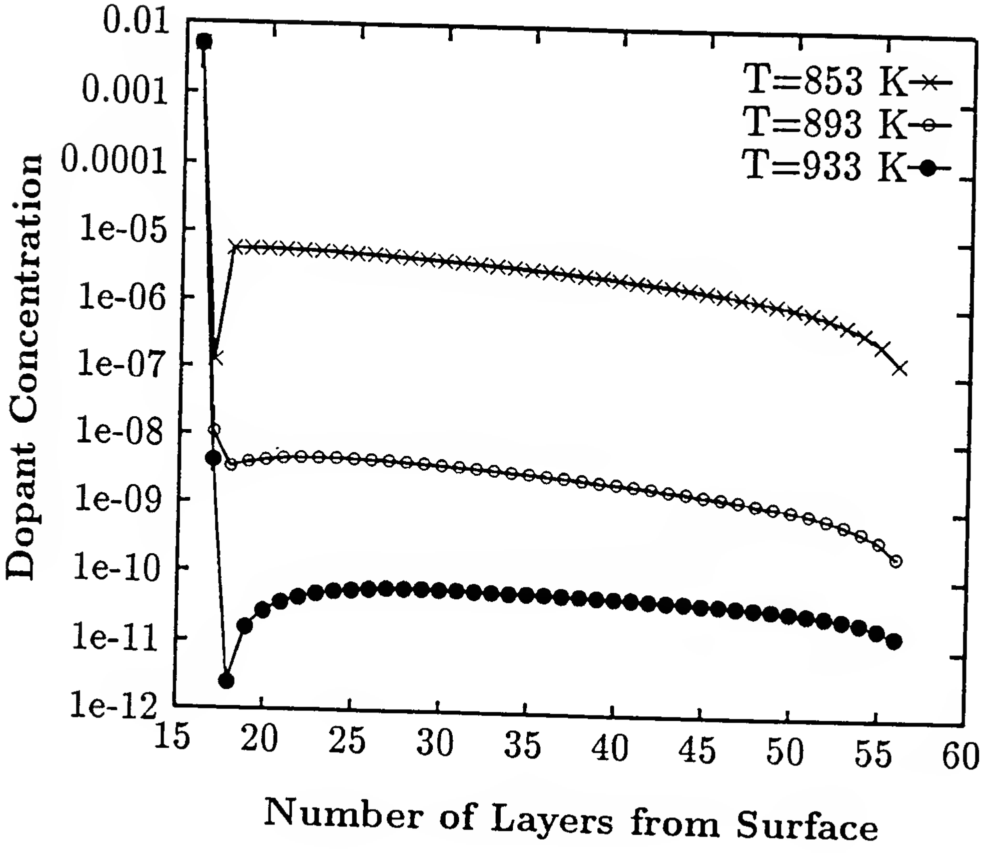


Fig. 9. Plots of dopant segregation profiles for various growth temperatures for the growth time of 20 seconds.

# PERFORMANCE EVALUATION AND IMPROVEMENT OF A RESONANT DC LINK INVERTER WITH A LIMITED Q-FACTOR

Subbaraya Yuvarajan
Associate Professor
Department of Electrical Engineering

North Dakota State University
North University Drive
Fargo, ND 58105

Final Report for:
Research Initiation Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington , D.C.

and
North Dakota State University

December 1993

# PERFORMANCE EVALUATION AND IMPROVEMENT OF A RESONANT DC LINK INVERTER WITH A LIMITED Q-FACTOR

Subbaraya Yuvarajan
Associate Professor
Department of Electrical Engineering
North Dakota State University

## Abstract

Resonant DC Link (RDCL) inverters have several advantages compared to conventional dc link inverters. Some of the problems associated with an RDCL inverter with a limited Q-factor are studied. An experimental inverter was built mainly to make quantitative measurements and analyze problems like zero-crossing failure. The losses occurring in the resonant link were obtained for different link-circuit parameters. A digital storage oscilloscope and digital data processing software were used in the computation of the losses. The complete control circuit of the inverter system incorporating Sine PWM control for the inverter switches was developed. The control circuit which was developed for a single-phase inverter was extended to a three-phase inverter.

The zero-crossing failure in an RDCL inverter is a serious problem which reduces the efficiency of the inverter system and increases the power rating of the link-shorting switch. First the effect of load on the minimum link voltage was studied. A simple method of using current-feedback for eliminating the zero-voltage crossing problem was first established using computer simulation. The method was implemented on an experimental inverter system by incorporating current-feedback by using current-sensing Power MOSFETs as inverter-switches. The details of the complete control circuit incorporating current feedback and the experimental waveforms are presented.

# PERFORMANCE EVALUATION AND IMPROVEMENT OF A RESONANT DC LINK INVERTER WITH A LIMITED Q-FACTOR

Subbaraya Yuvarajan

## 1. INTRODUCTION

High-efficiency power converters are very useful in aircraft and space applications. Resonant DC Link (RDCL) inverters introduced recently help to keep the switching loss negligibly small even at high switching frequencies [1] - [4]. The addition of an L-C resonant section in a conventional inverter results in an oscillating link with periodic zero-crossing points. If the inverter switches are turned-on and off at the zero-voltage points, there will not be any power loss in the switches. The only condition is that the operation of the inverter switches is to be synchronized to the oscillatory voltage waveform of the resonant link. If the frequency of the resonant link is very high, the control signal of the inverter derived from considerations like PWM remains unaltered.

While the concept of RDCL is very attractive, there are some basic difficulties in maintaining a stable link with definite zero-crossings. The problem is mainly due to a low Q-factor of the L-C section. The link voltage of an RDCL with an ideal inductor and a capacitor goes through zero voltage at the end of each resonant cycle. When the inductor has a parasitic series-resistance, the link voltage gets damped out. A switch connected across the capacitor is closed for a short time at the end of the resonant cycle so as to supply some energy to the inductor to offset the damping effect. Fig. 1 shows the power circuit of an RDCL feeding a single-phase inverter and the block diagram of the control circuit. The link inductor $L_C$, the capacitor $C_C$, and the link-shorting MOSFET QL constitute the link section. The link-shorting Power MOSFET is driven by a clock signal whose frequency is adjusted to be equal to the resonant frequency of the link. It is enough if the duty cycle of the waveform is very small.

The project is to study the performance of an RDCL inverter with a resonant link having a limited Q-factor. Fig. 2 shows the parasitic elements in the link section and the shape of the capacitor voltage in a typical RDCL. The minimum capacitor voltage $V_{min}$ is a critical parameter which decides the efficiency of the inverter. The effect of varying the values of the link elements on the shape of the link

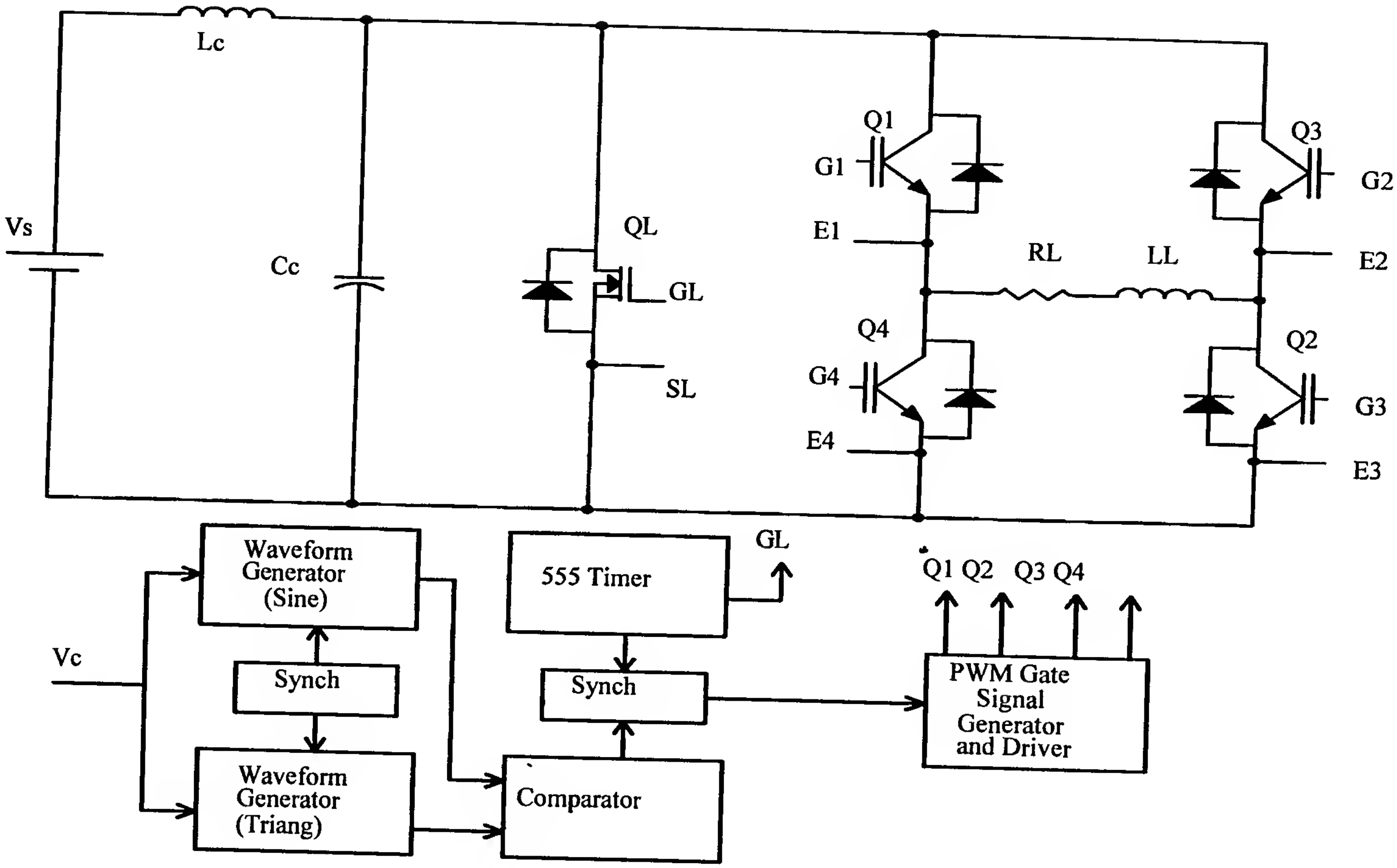

Lc
Vs
Cc
QL
GL
SL
Q1
G1
E1
Q4
G4
E4
RL
LL
Q3
G2
E2
Q2
G3
E3
Waveform Generator (Sine)
Synch
Waveform Generator (Triang)
Vc
555 Timer
GL
Synch
Comparator
Q1 Q2 Q3 Q4
PWM Gate Signal Generator and Driver

Fig. 1 Power circuit of an RDCL inverter and block diagram of control circuit

voltage and the power loss in the link will be studied. In particular, the variation of $V_{min}$ with the values of $L_c$ and $C_c$ will be studied. The project also proposes a method to reduce the value of $V_{min}$ and to ensure zero-voltage crossing. Current Sensing MOSFETs will be used to provide current-feedback which will be used to adjust the frequency and duty cycle of the gate signal to QL. The method will be first verified by simulating the proposed control scheme using PSPICE [5]. It will be implemented on a single phase RDCL inverter.

## 2. THE RESONANT DC LINK

In analyzing the operation of the RDCL, the inverter is modeled as a constant current source [6]. Fig. 3 shows the resulting circuit diagram. If the elements are ideal, R = 0. With the switch S open, the capacitor voltage $v_c(t)$ is given by

$$v_c(t) = V_s(1 - Cos\omega_r t) \tag{1}$$

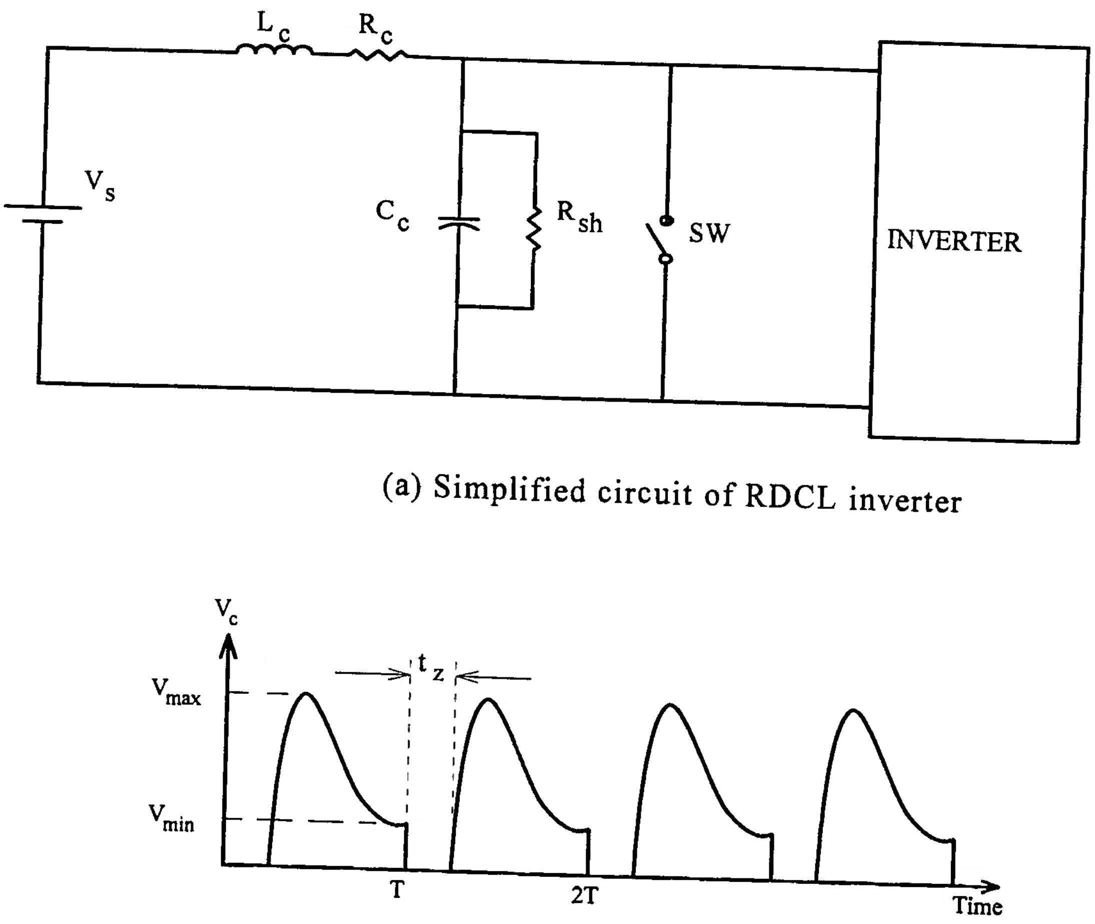
where $\omega_r = \dfrac{1}{\sqrt{LC}}$.

If the parasitic resistance of the inductor is included, the voltage across the capacitor is damped out. The minimum voltage across the capacitor $V_{min}$ is given by [6]

$$V_{\min} = (V_s - I_d R)[1 - \frac{1}{\sqrt{1-\xi^2}} e^{-\alpha t_2} Sin(\omega t_2 + \theta) + \frac{(I_m - I_d)Z}{\sqrt{1-\xi^2}} e^{-\alpha t_2} Sin\,\omega t_2 . \tag{2}$$

It is seen from (2) that zero-voltage crossing is lost if $V_{min}$>0. The value of $V_{min}$ depends essentially on the value of ($I_m$-$I_d$) and the values of L and C.

In the experimental inverter built, a Power MOSFET is used to realize the link-shorting switch. The block diagram of the link-switch signal generator is shown in Fig. 4. The width and frequency of the gate pulse to QL is varied by varying the potentiometers Rw and Rf respectively. The TTL output from the 555 Timer is converted into a 15 V signal using a high-speed driver SG3626. The frequency of the gate signal to QL is around 48 kHz.



(a) Simplified circuit of RDCL inverter

(b) Waveform of link voltage

Fig. 2. RDCL inverter with a limited Q-factor.

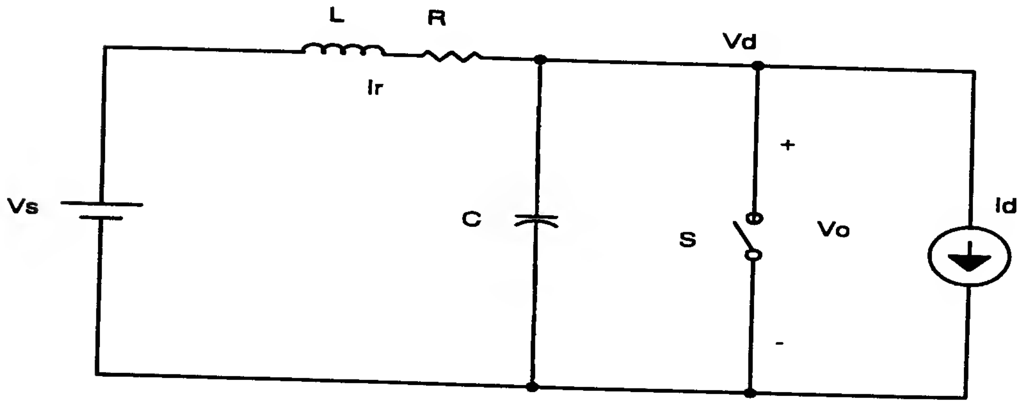


Fig. 3 Circuit model of an RDCL with inverter load

Several designs of the link inductor were tried. In one design, the inductor was built using a well-insulated flat-copper conductor. In another design, a multi-strand litz wire was used [7]. In both the designs, only air-cored inductors were built. The inductance-value was varied by changing the number of turns. The coil resistance was reduced by using parallel construction. The waveforms of the capacitor voltage and the inductor current under no load were recorded for several $L_cC_c$ combinations. The power loss in the link section was computed by recording the waveforms of the current and voltage associated with all the link elements. It was found that the value of $V_{min}$ was negligibly small and the power loss in the link section was less than 3 W for a dc input voltage $V_s = 30$ V and for different $L_cC_c$ combinations.

## 3. SINE PWM CONTROL FOR RDCL INVERTER

There are different type control schemes available for controlling the magnitude of the inverter output voltage and reducing the harmonics in the output. The most common one is the Sine Pulse Width Modulation (SPWM) [8]. The main advantages of this scheme are: reduced harmonics and the ability to control the voltage and frequency in the same power circuit. If P is the number pulses per half cycle of the reference sine wave, then all harmonics below (2P - 1) will be eliminated from the output. With the lower order harmonics eliminated, one needs to use a smaller filter inductance in the load circuit. The synchronization between the gate signal of the link switch and the SPWM control signal should be considered in the design of the RDCL inverter to ensure zero-voltage switching. To achieve this, a TTL monostable multivibrator and a flip-flop can be used.

In the hardware realization of SPWM, one has to generate a variable-frequency sine wave and a synchronized triangular wave while maintaining the ratio $f_t/f_s$ between the two frequencies (or the value of P) constant. Generation of a distortion-free sine wave and the synchronization between the sine and triangular waves are the main problems to be solved in the implementation. In the proposed control circuit, both the sine and triangular waveforms are generated using high-precision waveform generating ICs like ICL 8038 [9]. Fig. 5 shows the block diagram of the synchronized sine and triangular waveform

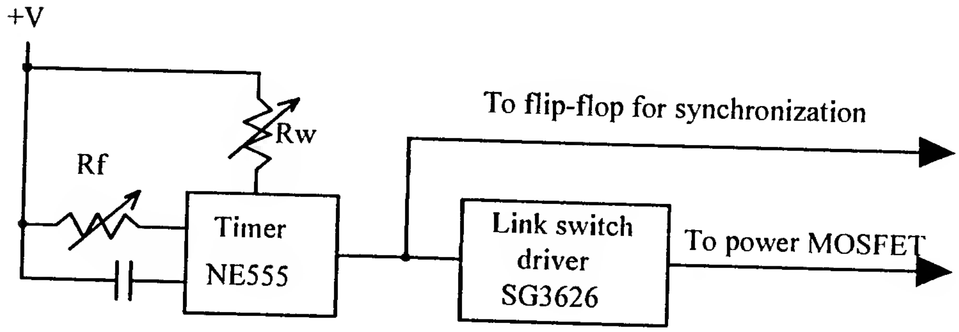


Fig. 4  Block diagram of link-switch gate signal generator

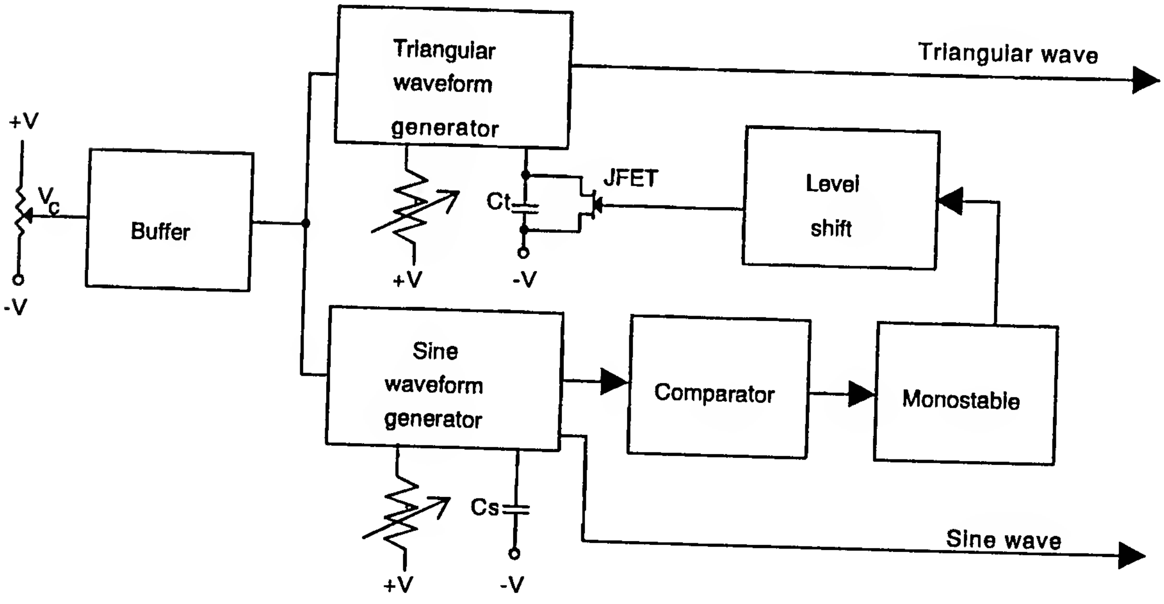


Fig. 5  Block diagram of sine-triangular waveform generator

generators. The frequencies of the two waveform generators can be continuously varied by a single control input $V_C$. The two waveforms are synchronized with the help of an external initialization circuit.

Fig. 6 shows the Complete SPWM control circuit of the proposed RDCL inverter. The number of pulses per half cycle, P, is varied by adjusting the timing resistors and capacitors ($R_tC_t$ or $R_SC_S$). The sine and triangular waveforms are synchronized by connecting a JFET across the timing capacitor ($C_t$) of the IC which generates the triangular waveform. A narrow pulse derived from the sine-wave signal triggers the JFET, thereby shorting and initializing the capacitor $C_t$. A comparator (OA2) and a monostable provide the gate pulse to the JFET. If the width of the gate pulse applied to the JFET is small and if $f_t/f_S$ is adjusted to be an integer, then the triangular waveform will be continuous. The ratio of frequencies can be adjusted to be an integer by adjusting the potentiometers $R_t$ or $R_S$.

The part of the circuit that generates the gate signals for the inverter-switches from the sine and triangular signals is shown in Fig. 7. The OP Amp U7 connected as a non-inverting amplifier controls the amplitude of the sine wave and hence the modulation index m. The comparator U11A compares the sine and triangular waves. The output from U11A is passed through a set of inverters and NAND gates to obtain the SPWM gate signals for the inverter-switches. The SPWM gate signal is synchronized to the link-switch gate signal with the help of a flip-flop U12. The synchronization ensures zero-voltage switching of the IGBTs in the inverter.

The experimental waveforms generated by the control circuit are recorded using a digital oscilloscope. Fig. 8 shows the sine and triangular waveforms and the narrow gate pulses applied to the JFET for an output frequency of 60 Hz and a frequency ratio $f_t/f_S = 12$. Fig. 9 shows the response of the waveform generators for a step change in $V_C$. It can be seen that the response is fast and the ratio of frequencies remains constant. The output of comparator U11A and the gate pulses to the inverter switches are shown in Fig. 10.

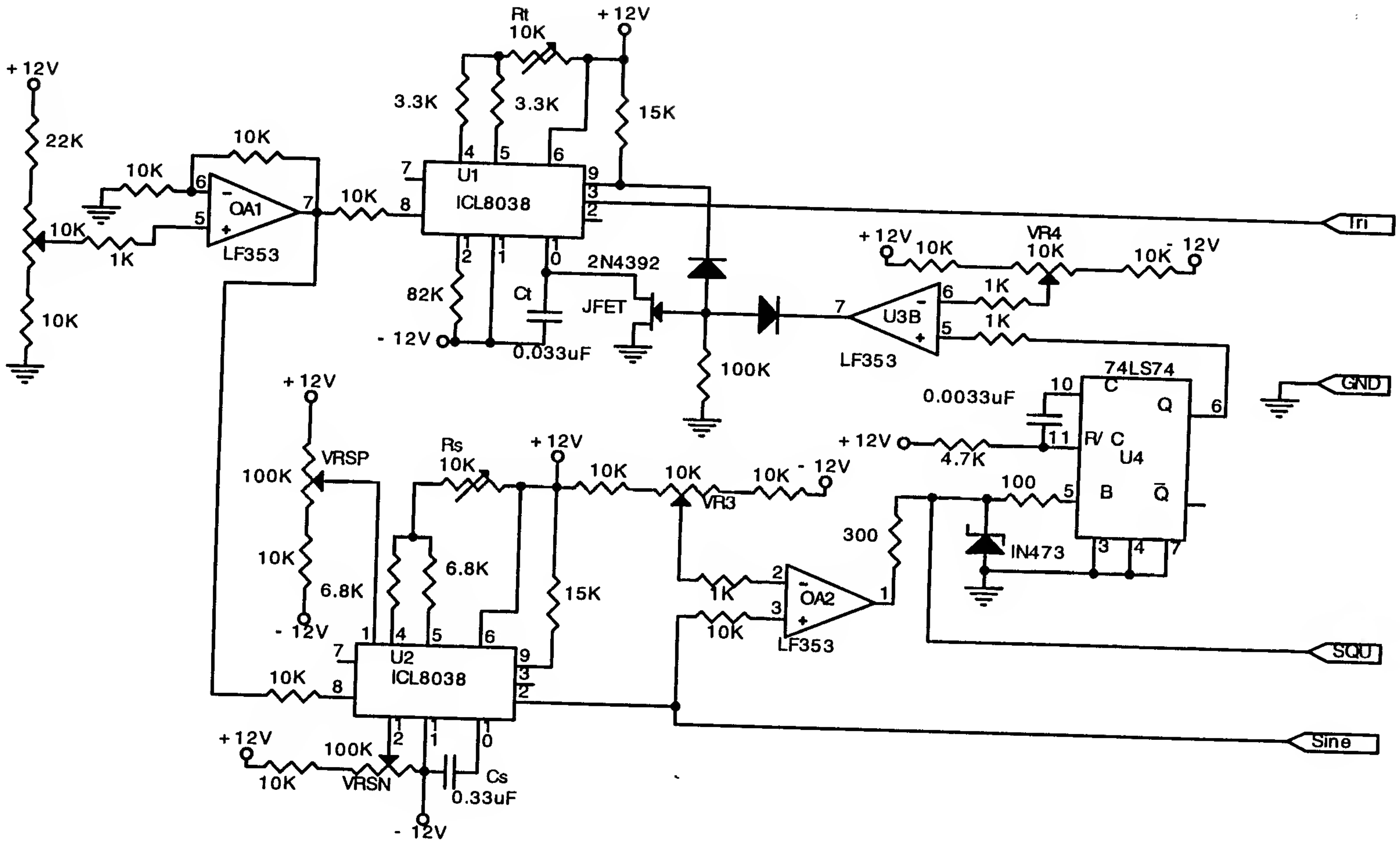


Fig. 6 Complete circuit of sine-triangular waveform generator

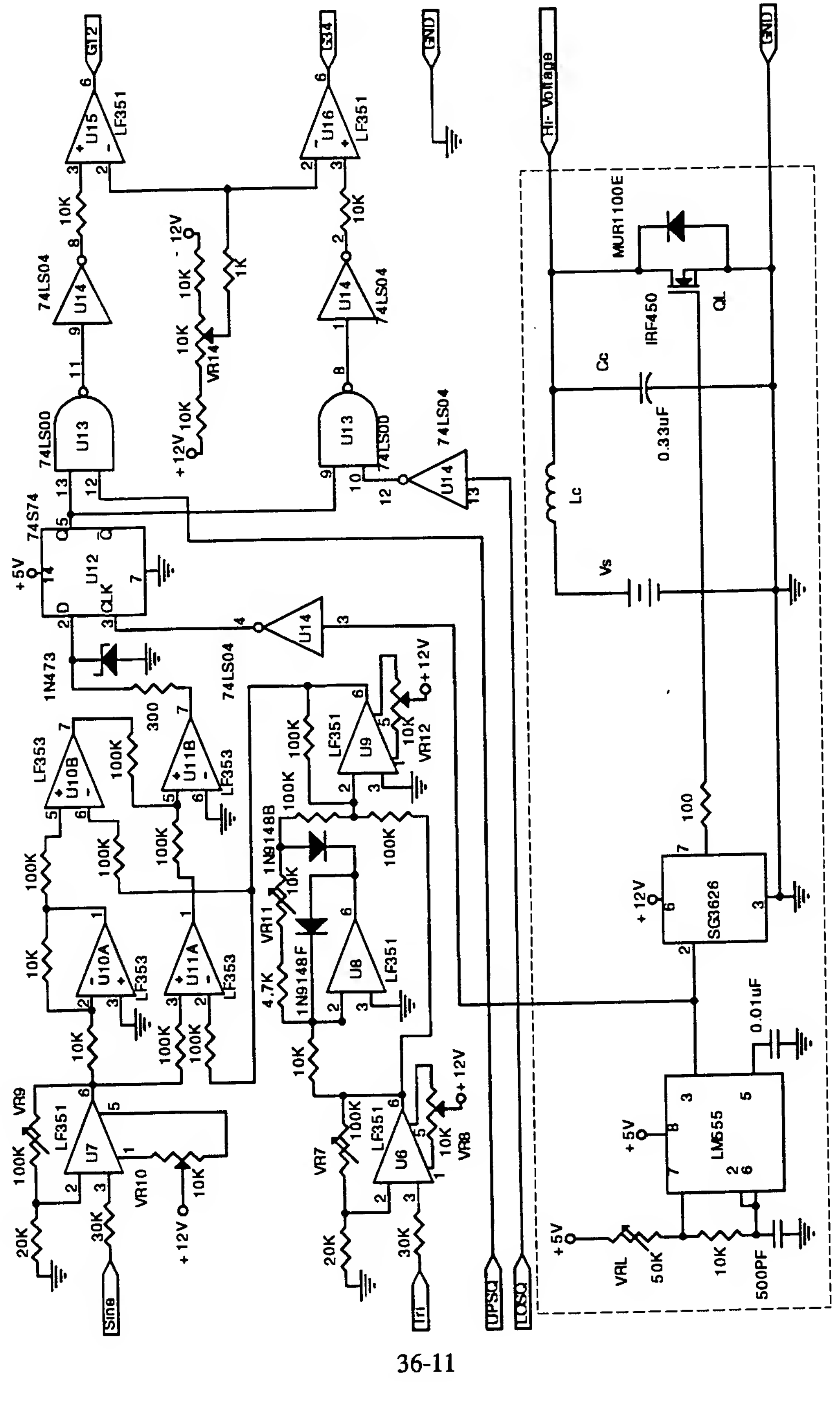
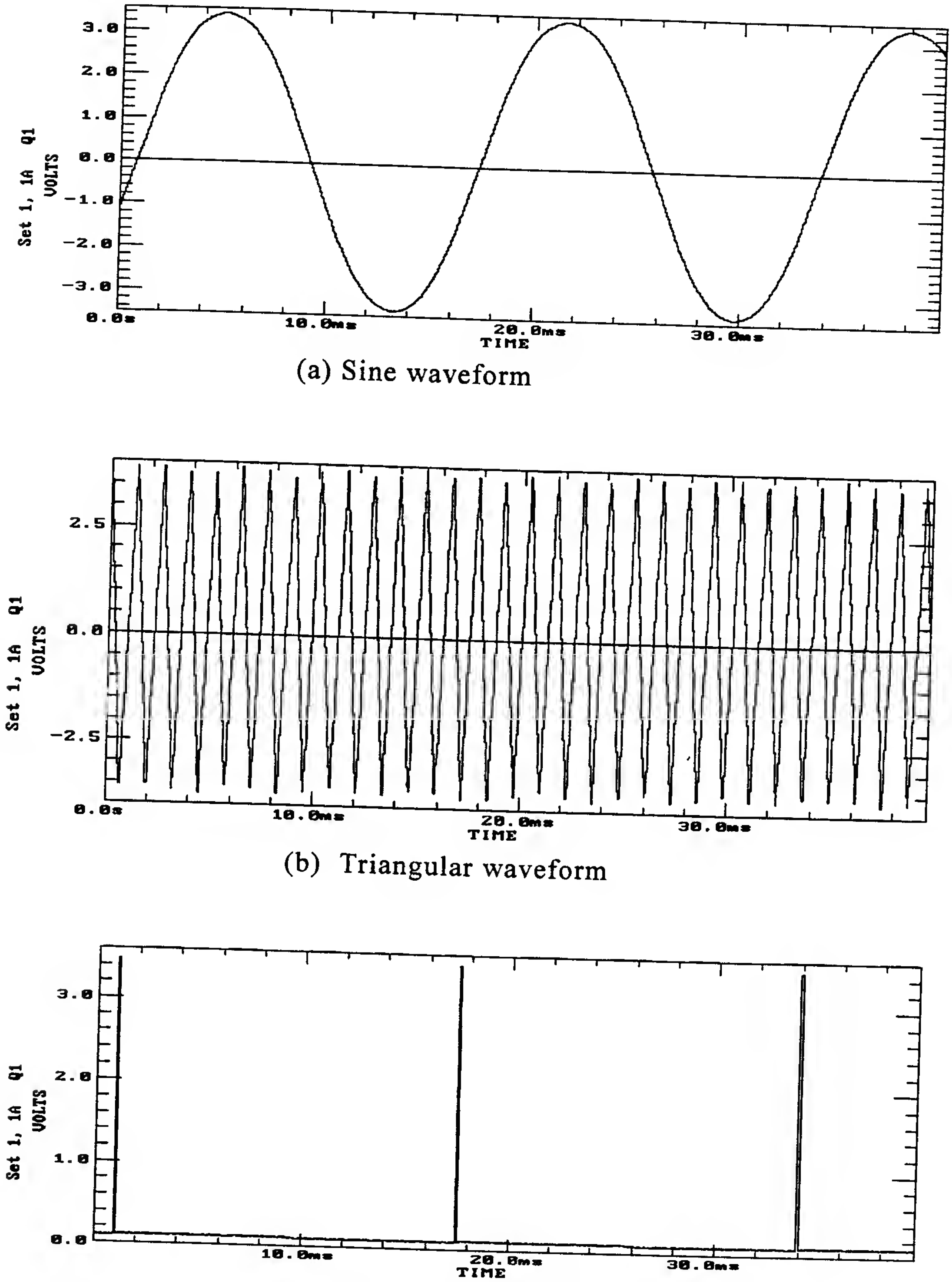


Fig. 7 Complete circuit of PWM signal generator

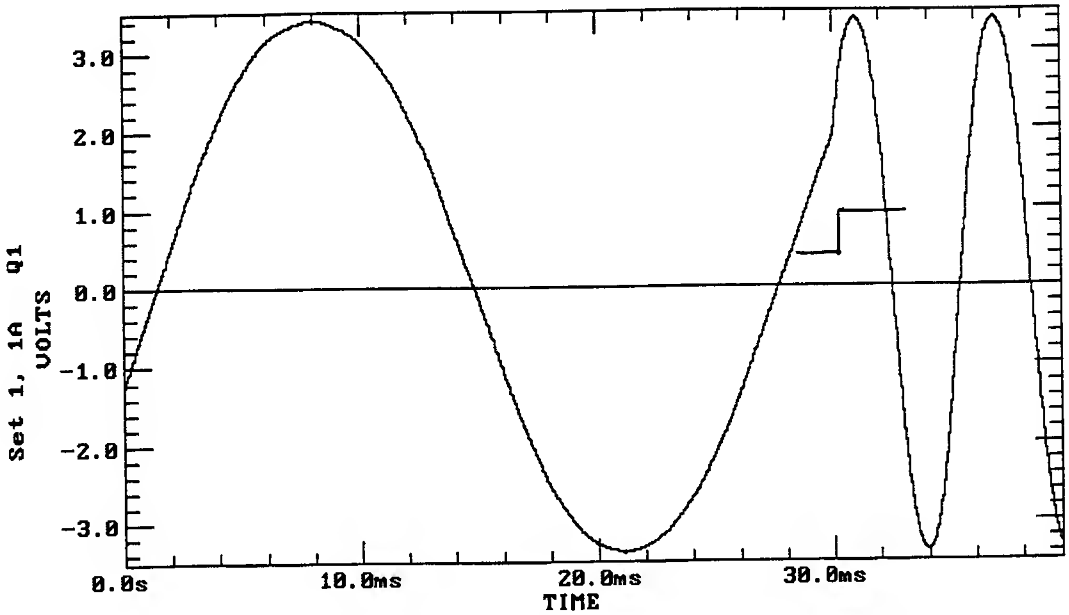


(a) Sine waveform

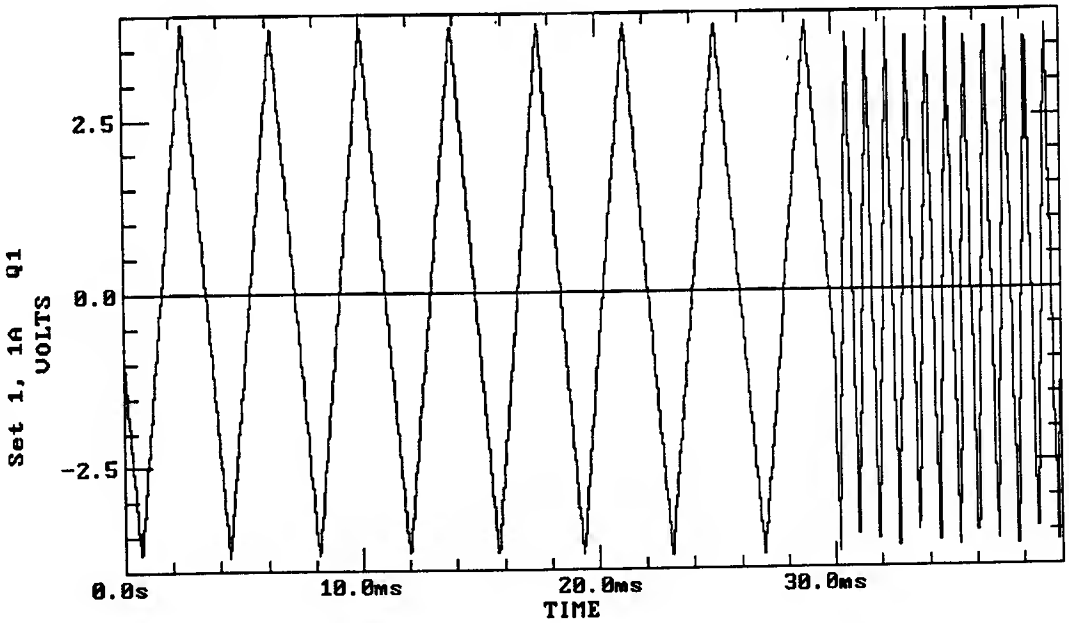
(b) Triangular waveform

(c) Gate voltage for JFET

Fig.8 Experimental sine, triangular, and gate pulse waveforms

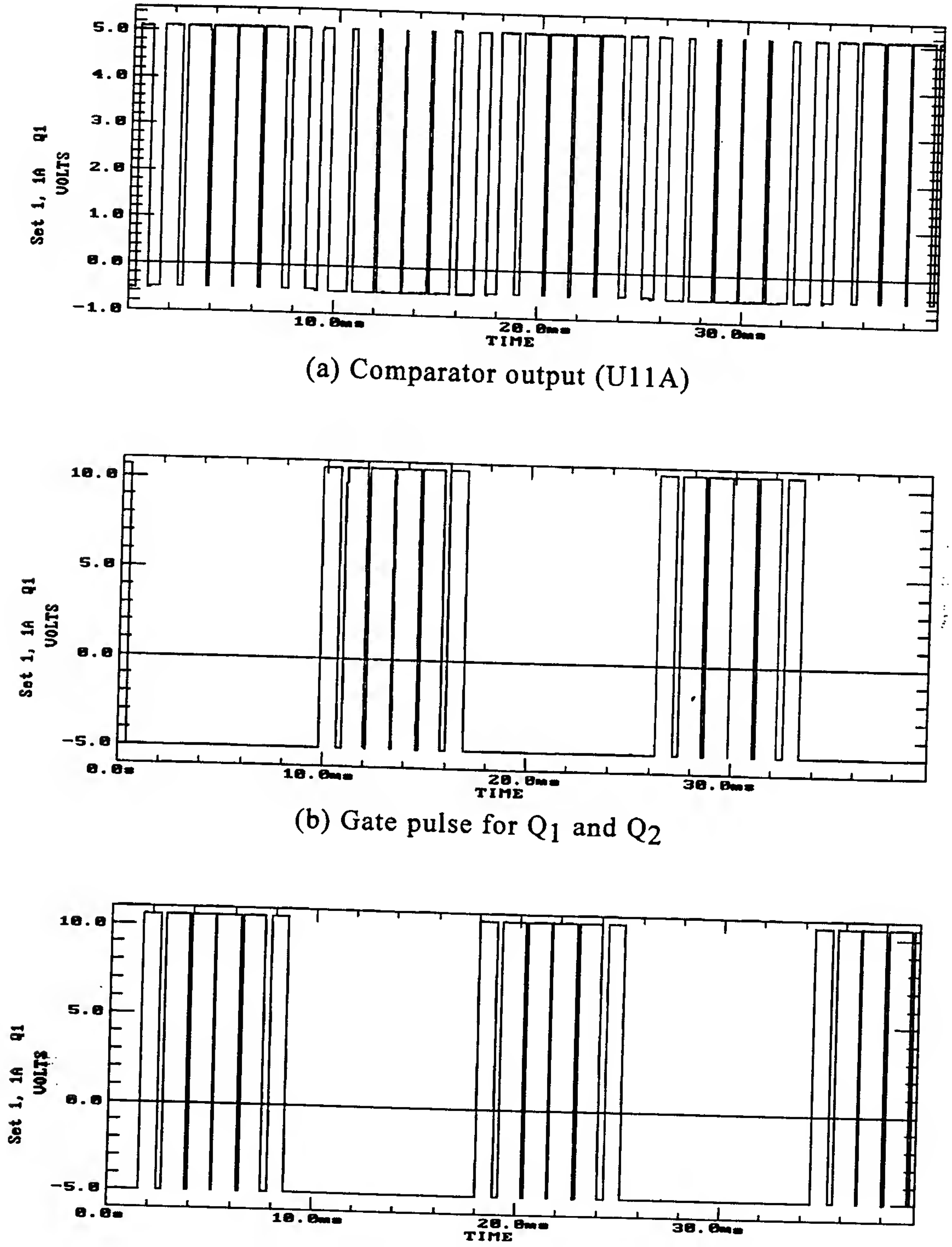


(a)Sine waveform



(b)Triangular waveform

Fig. 9 Step response of waveform generators



(a) Comparator output (U11A)

(b) Gate pulse for $Q_1$ and $Q_2$

(c) Gate pulse for $Q_3$ and $Q_4$

Fig. 10 Output of comparator and gate pulses to inverter switches

## 4. POWER CIRCUIT OF RDCL INVERTER

The power circuit of a single phase RDCL inverter is shown in Fig. 11. That gate signals for the IGBTs are obtained using dedicated driver EXB 851 supplied Fuji Electric [10]. Two EXB 851 drivers supply the gates of the high-side IGBTs Q1 and Q3. Two push-pull drivers supply the gates of the two low-side IGBTs Q2 and Q4. The driver EXB 851 provides the isolation necessary for the high-side switches. It also protects the IGBT against over-current by feeding back the collector voltage through a fast-recovery diode MUR 1100E. The driver circuit and the power circuit are shown in Fig. 11.

The waveforms of voltages and currents in the RDCL inverter were recorded using a digital storage oscilloscope. Fig. 12 shows the waveforms of the gate signal to the link-switching Power MOSFET and the link capacitor voltage. With inductive load, there are definite zero-crossings. However, with resistive load, the capacitor voltage fails to go to zero and the MOSFET QL is forced to switch at a finite voltage, $V_{min}$. The minimum voltage could be as high as 15 V. A possible solution to the above problem is described in Section 8.

The load voltage and load current waveforms obtained on the RDCL inverter for an inductive load are shown in Fig. 13. The Sine PWM scheme uses a reference frequency of 60 Hz, P = 6, and a modulation index of 0.9. The harmonic content in the output is obtained by using a digital data processing software. A Fast Fourier Transform (FFT) performed on the experimental waveform of voltage (or current) yields the frequency spectrum a shown in Fig. 14. It is seen that the output is free from several lower order harmonics below (2P - 1). The spectrum shows a harmonic component around the resonant frequency of the link which can be easily filtered.

## 5. THREE-PHASE RDCL INVERTER

Three-phase RDCL inverter is suitable for driving induction motors. It uses the resonant dc link described in Section 2. The Sine PWM control first developed for a single-phase inverter is extended to a three-phase system. The block diagram of the three-phase sine wave generator is shown in Fig. 15. The sine wave signal of Phase-A and a synchronized triangular wave are generated using two ICL 8038 waveform generators as described in Section 3. In order to get a balanced set of waveforms, the value of P

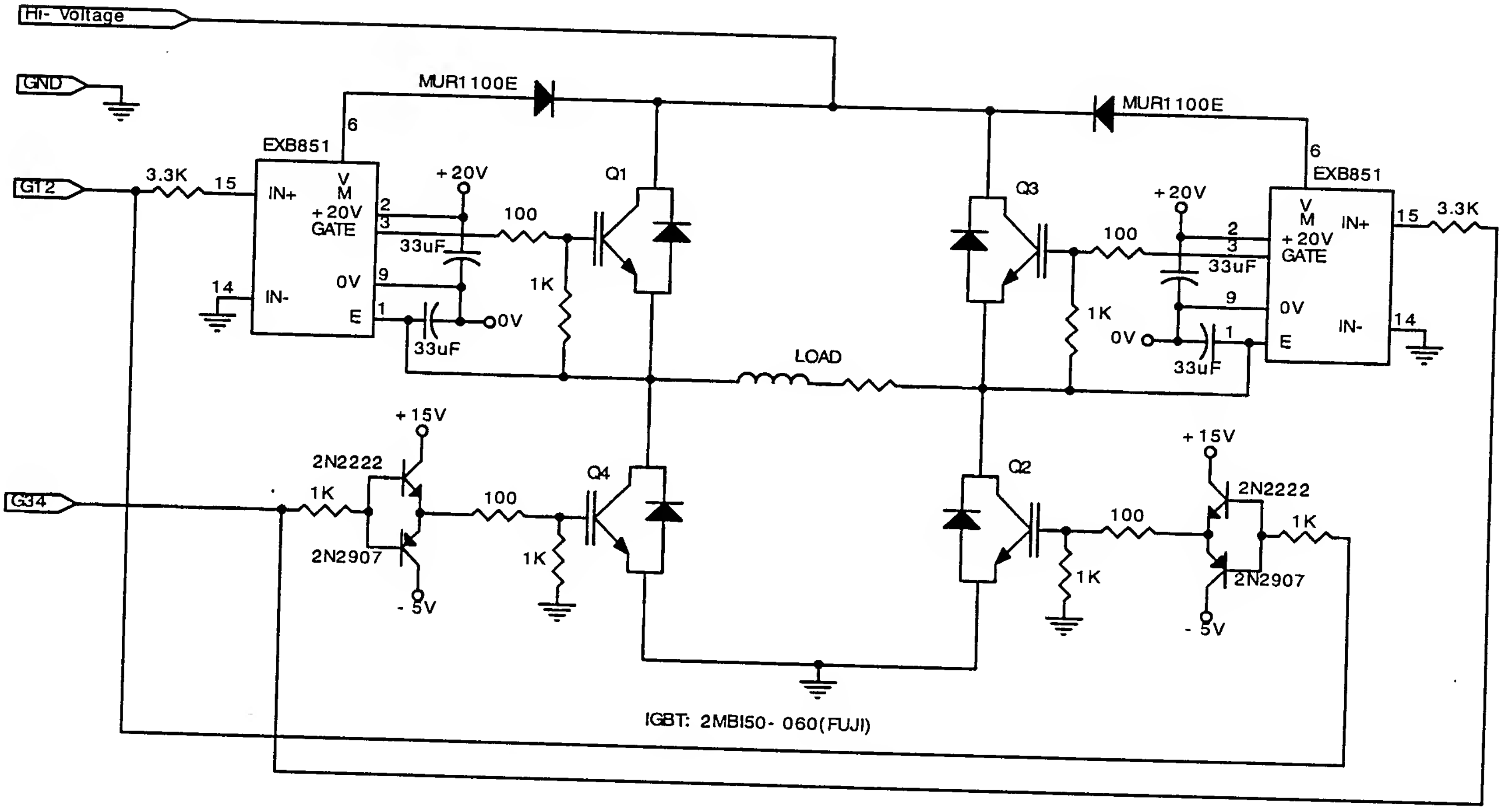
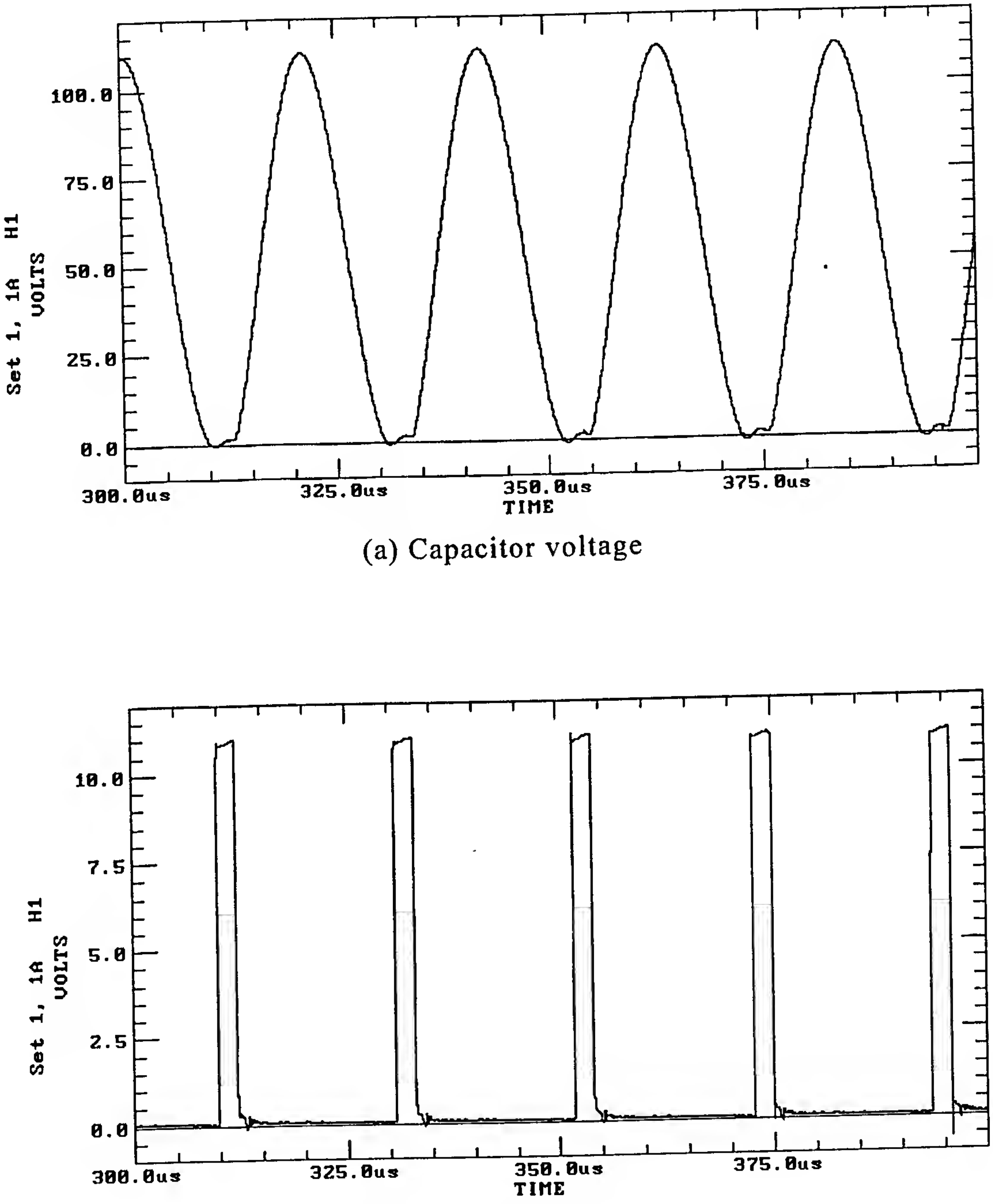


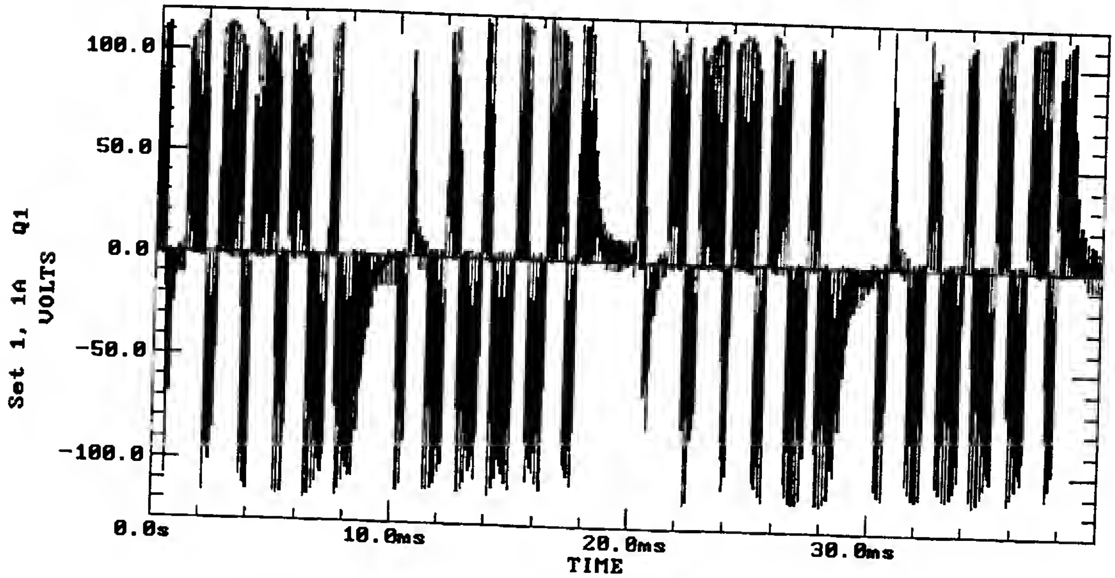
Fig. 11 Driver circuit and power circuit of RDCL inverter



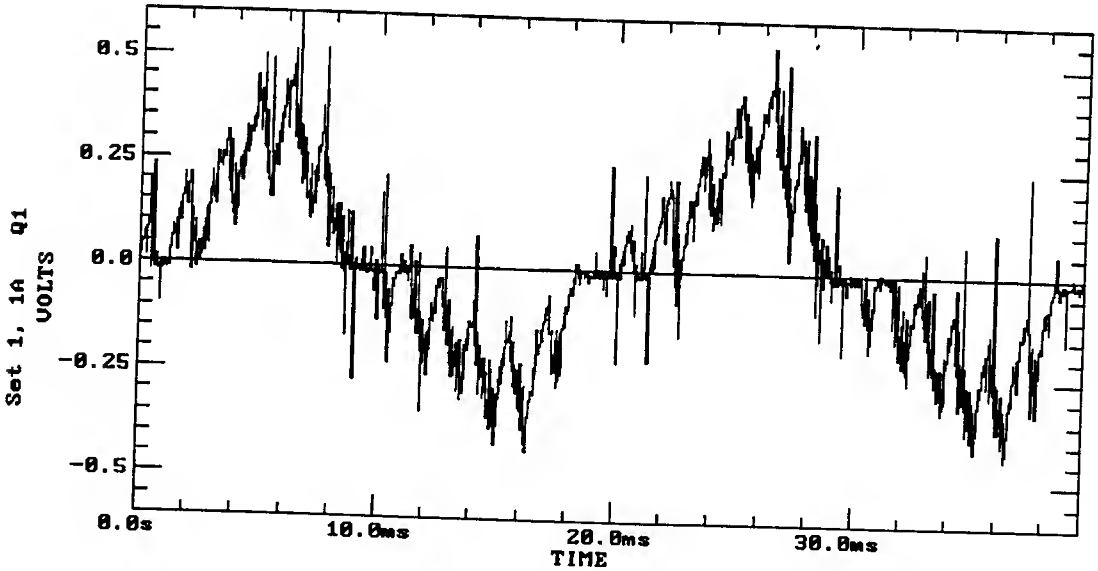
(a) Capacitor voltage

(b) Link-switch drive signal

Fig. 12 Experimental waveforms of link signals

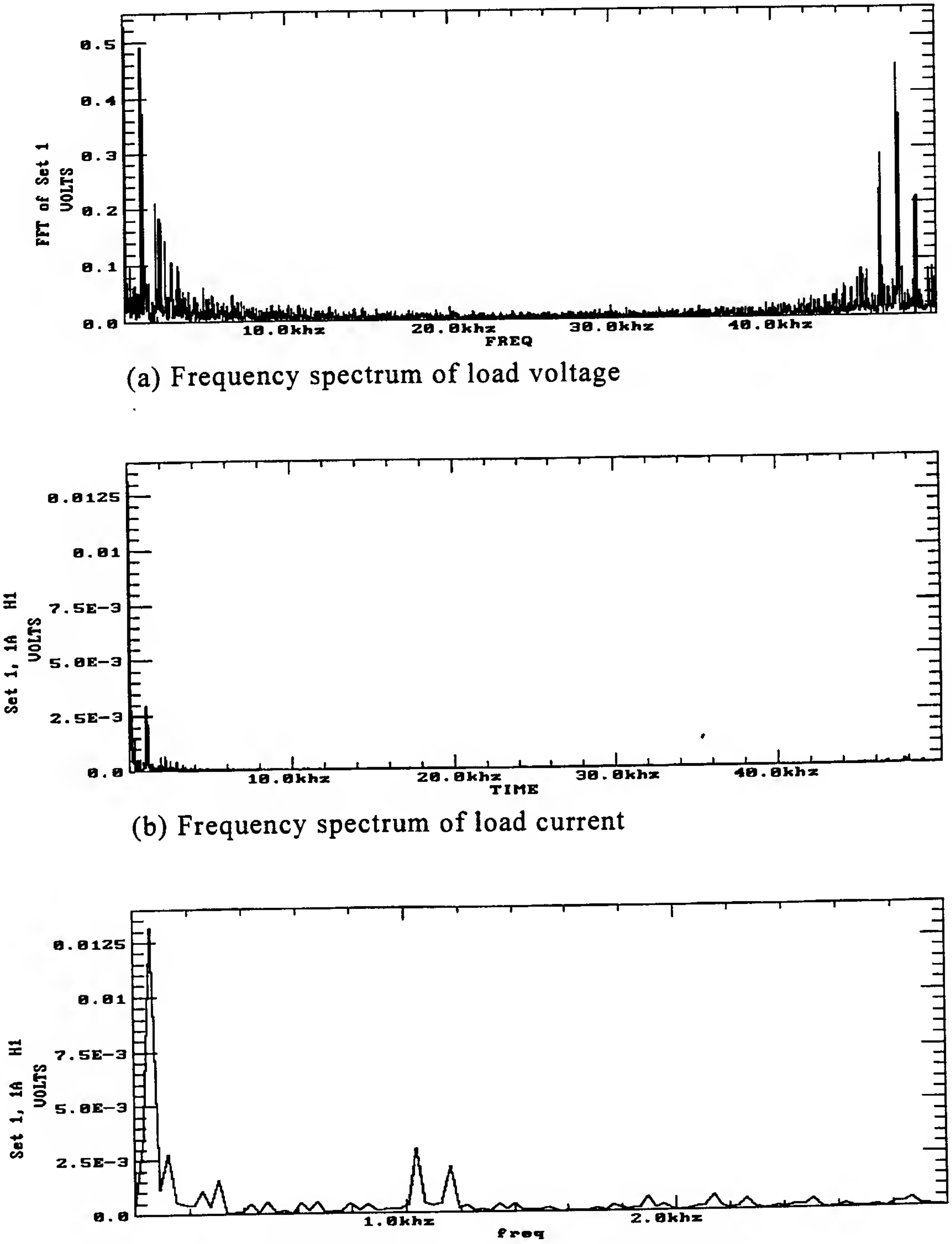


(a) Load voltage



(b) Load current

Fig. 13 Load waveforms of RDCL inverter



(a) Frequency spectrum of load voltage

(b) Frequency spectrum of load current

(c) Expanded frequency spectrum of load current

Fig. 14 Frequency spectrums of load voltage and current

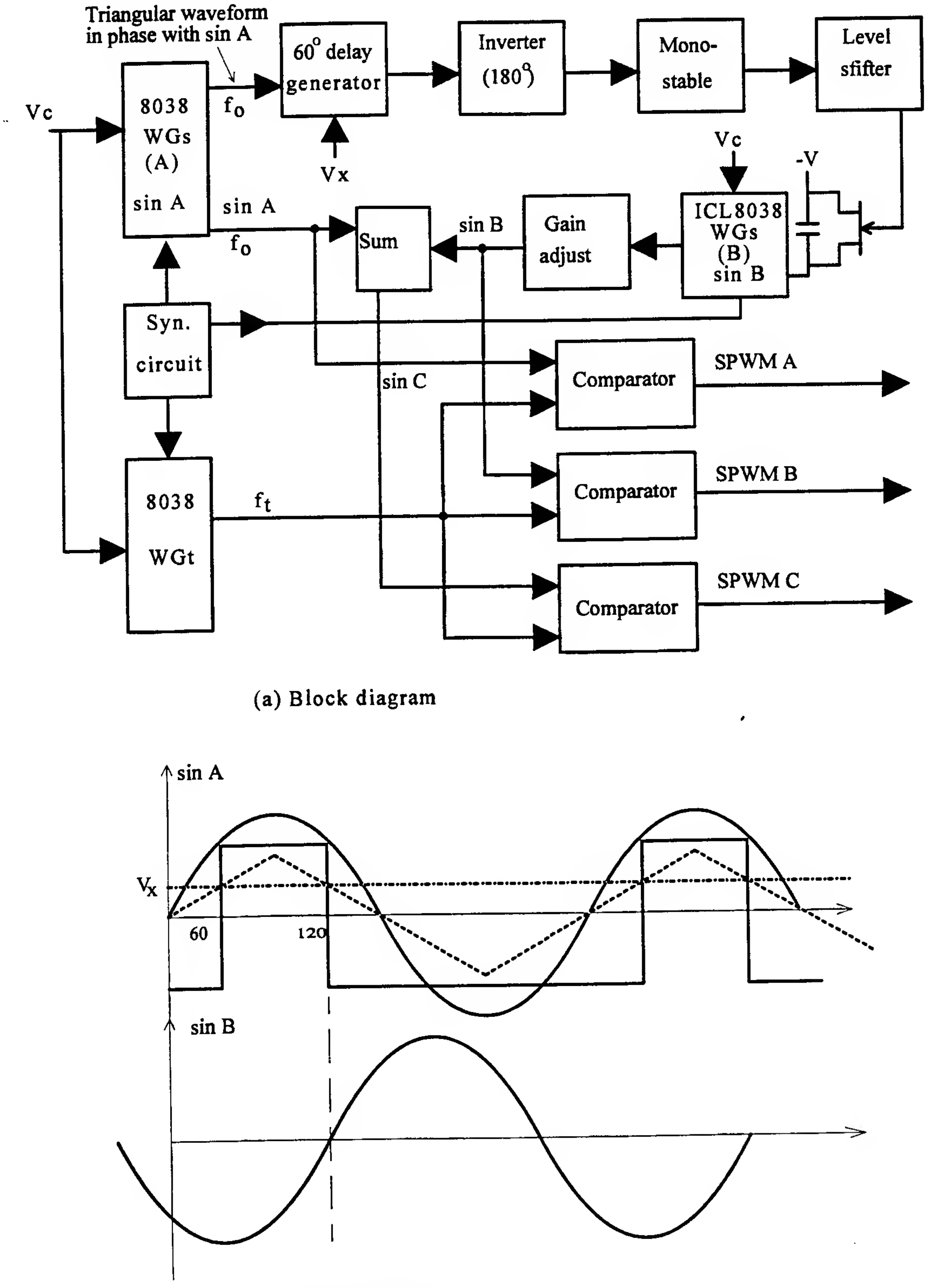
should be a multiple of three. The Phase-B sine wave is generated using a third waveform generator and is synchronized to the Phase-A sine wave. The synchronizing pulse is derived from the Phase-A triangular wave as shown in Fig. 15. The phase shift of 120 deg is obtained by comparing the Phase-A triangular wave and a reference $V_X$. The sine wave for Phase-C is generated from the sine waves of Phase-A and Phase-B using the relation SinC = -(SinA+SinB). Fig. 16 shows the triangular wave and the sine waves of Phases A and B.

The power circuit of the inverter was built using six IGBTs available as a six-pack module. The control circuit uses three EXB 851 drivers for the three high-side switches and three push-pull drivers for the low-side switches. A three-phase induction motor was supplied by the RDCL inverter. Fig . 17 shows the load voltage and load current waveforms of the induction motor.

## 6. PERFORMANCE OF AN RDCL INVERTER

While the concept of RDCL inverter is very attractive, there are some basic problems caused by the low value of Q-factor of the resonant circuit. The link voltage of an RDCL inverter with an ideal inductor and capacitor goes through zero at the end of each resonant cycle. The L-C section in a practical inverter has a finite Q-factor whose value is limited by the parasitic series resistance of the inductor and the leakage resistance of the capacitor shown in Fig. 2. Consequently, the link voltage does not go to zero and the link-shorting Power MOSFET is to be turned-on at a non-zero voltage, resulting in additional power loss. The minimum link voltage ($V_{min}$) and the power loss in the link elements are found to vary with the power factor of the load, the resonant frequency, and the zero-voltage interval ($t_Z$). The above observations were made on the experimental inverter described in Section 4.

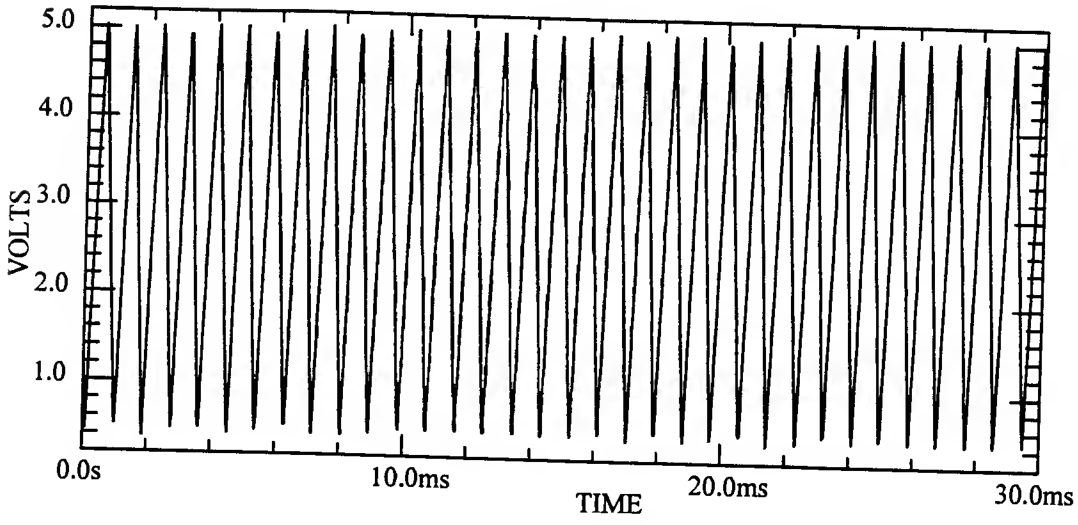
This section investigates some of the problems associated with an inverter having a limited Q-factor. The performance of the inverter was studied by recording all the waveforms of voltage and current associated with the link and measuring the power loss using a storage oscilloscope and a digital data processing software. Certain modifications in the link operation that will improve the overall performance of the inverter are being proposed in a later section.



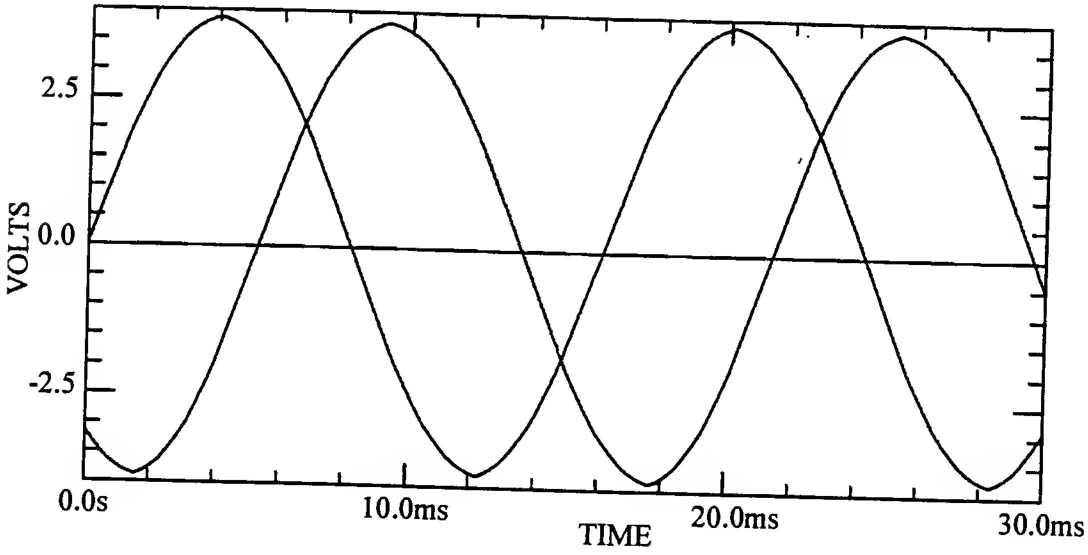
(a) Block diagram

(b) Illustrative waveforms

Fig. 15 Block diagram and illustrative waveforms of 3-phase sine wave generator

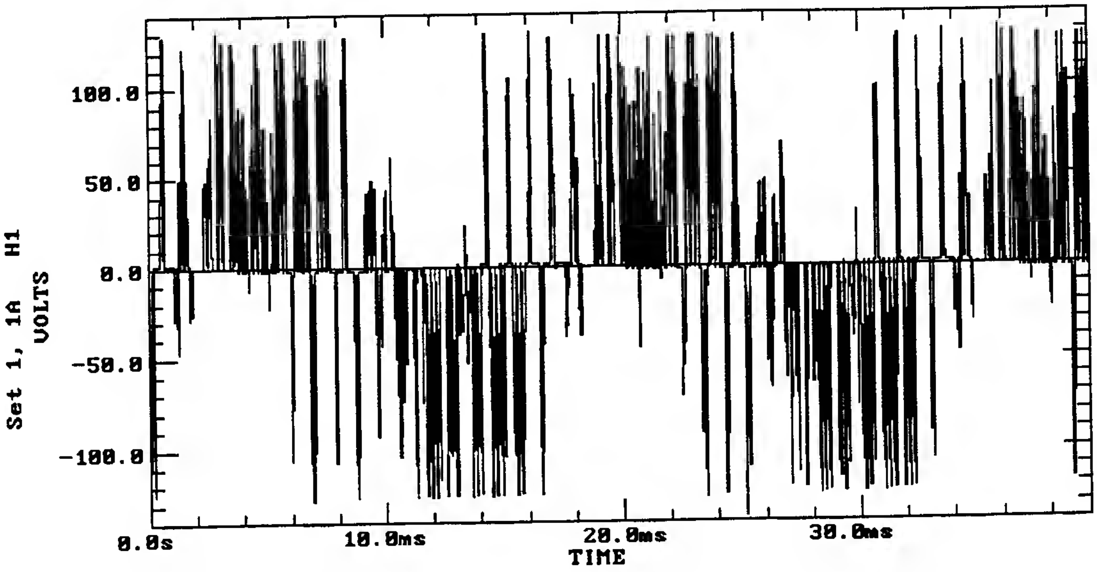


(a) Triangular waveform

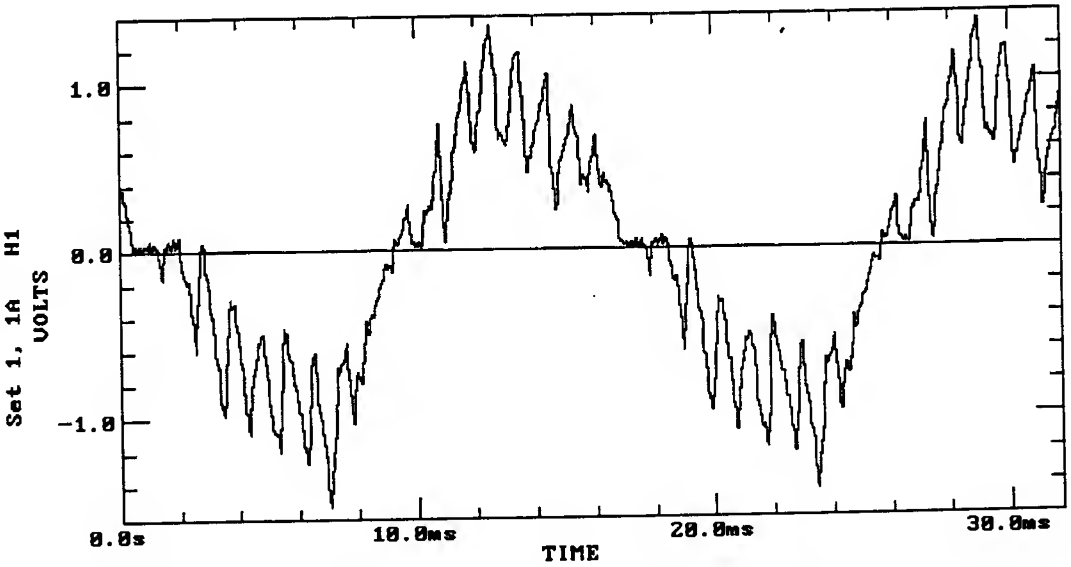


(b) Sine waveforms of phase A and phase B

Fig. 16 Experimental waveforms of 3-phase sine wave generator



(a) Load voltage



(b) Load current

Fig. 17 Waveforms of 3-phase RDCL inverter feeding an induction motor

a) Effect of low Q-factor

A resonant dc link with a high value of Q-factor is difficult to design. The RDCL built in the laboratory has Q of 45 at the link frequency of 48 kHz. The resonant frequency is chosen high so that the pulse density modulation used in the inverter does not introduce any appreciable error. A high resonant frequency means a higher power loss due to the frequent closing and opening of the link-switch at a non-zero voltage. A non-zero link voltage requires the use of a separate link-switch instead of the inverter switches serving the purpose. In addition, the link-shorting MOSFET has to be rated for a high current and protected by a proper snubber and a heat sink. An estimation of the power loss in the switch will enable the choice of the link-shorting Power MOSFET and the design of the snubber.

The resonant frequency of the link is determined by the product $L_C C_C$. The maximum current through $L_C$ is determined by the ratio $C_C/L_C$. Since the link current determines the loss in the link section, a careful choice of the values of $L_C$ and $C_C$ has to be made in order to minimize the loss at a given resonant frequency. Since the inverter constitutes a nonlinear current source, computer simulation will be useful in analyzing its performance.

b) Measurement power loss in an RDCL inverter

The power loss in an RDCL inverter with a reasonably high value of Q will be very small. An accurate measurement of the power loss will enable the estimation of the efficiency of the inverter system and will also enable one to investigate methods of improving its efficiency. Power loss in the RDCL occurs in the link inductor, capacitor, and the link switch. An accurate measurement of the power lost in any circuit element can be done by first recording the waveforms of voltage and current associated with that element and then processing waveforms. The digital data processing software VU-POINT [11] was used to multiply the two waveforms and compute the average power loss. This method of computing the power loss can be used to identify the regions of voltage/current waveforms where considerable power loss occurs and to eliminate the source to the extent possible.

The voltage and current waveforms of the elements in the link contain high frequency transitions. The waveforms also include ac and dc components. Most of the loads are operated at 60 Hz which corresponds to a period of 16.667 ms. The resonant frequency of the link is 48 kHz which corresponds to a period of 20.8 μs. Since the digital oscilloscope uses an analog-to-digital (A/D) converter to acquire the waveform data, the sampling rate has to be high enough to catch the high frequency transitions. At the same time, the sweep time (record length) has to be long enough to cover at least one cycle of the output waveform. To sample the data at the rate of 10 Mega samples/second (sampling interval = 100 ns) and store one cycle of output at 60 Hz in memory, the minimum size of memory required is

$$M_s = \frac{16.667ms}{100ns} = 166.67\,kBytes \tag{3}$$

The Nicolet Oscilloscope Model Pro 42 has a sampling rate of 20 Mega samples/second and a channel memory of 1 MBytes, which are more than sufficient to store the data.

The power loss in the copper-foil or litz-wire type inductor is relatively small due to the absence of skin effect. The power loss in the inductor is mainly due to the resistive effect of the copper itself. The resistance of a coil made of 99.99% pure copper is given by

$$R(\Omega) = \frac{1.724 \times 10^{-6} \times length(cm)}{thickness(cm) \times width\,(cm)}. \tag{4}$$

The parasitic resistance of the copper-foil type inductor used in the experiment was calculated to be 10.344 mΩ.

The losses in the link capacitor is due its leakage (shunt) resistance. A polypropylene capacitor is used to keep the losses small. The typical value of leakage resistance provided by the manufacturer is 42500 MΩ. The capacitor also has a series resistance of 0.11mΩ.

## 7. SIMULATION OF AN RDCL INVERTER WITH CURRENT FEEDBACK

The problem of zero-crossing failure is linked to the nature of the load current. The capacitor current that is responsible for the zero-crossing failure is equal to the difference between the inductor current and the load current. It is possible to adjust the frequency and duty cycle of the gate signal to QL, through a feedback of the load current, in order to overcome zero-crossing failure. The feasibility of the method was first tested through computer simulation. In the simulation, the inverter switches were modeled as ideal voltage controlled switches. The program used a resistive load as an example. Fig. 18 shows the simulated waveforms of capacitor voltage and the gate pulses to QL. The simulation was started at t = 0 with no load. The link-voltage waveform shows periodic zero-crossings. At t = 100 μs, a resistive load was added. With a fixed-frequency, and fixed-duty-cycle gate signal to QL, it is seen that the addition of the load results in zero-crossing failure.

The waveform of Fig. 19 corresponds to the situation where the frequency and duty cycle are varied as a function of the load current. It is seen that the link voltage exhibits definite zero-crossings. With a careful choice of the frequency and duty cycle, the link voltage was forced to go through zero at the end of each resonant cycle. Fig. 19 also shows the fast decay of the link voltage if there are no pulses applied to the link-shorting MOSFET (after 200 μs). The gate voltage to QL has to be modified based on the value of the load current and the instant of application of the load. Thus, one has to design the 555 Timer with a nominal value of frequency and duty cycle and vary them suitably whenever the load changes (in other words, the gate signals to the inverter switches change). The practical implementation of the current feedback is described in the next section.

## 8. RDCL WITH CURRENT FEEDBACK

One has to sense the load current before providing a feedback. Current sensing MOSFETs have been developed to meet the need of current monitoring in the power device [12]. The power circuit of the inverter was built using current sensing Power MOSFETs. Fig. 20 shows the power circuit of the single-phase inverter, the driver circuit, and the current sensing circuit. The dedicated driver for the Power MOSFET, IR 2110 was used in the circuit. A single driver supplies the gate signals for both the low and high-side switches in one arm. The sum of the drain currents of Q2 and Q4 gives the total load current. The sense outputs of the low-side MOSFETs Q2 and Q4 are converted into proportional voltages and added together to give the total signal $V_{st}$.

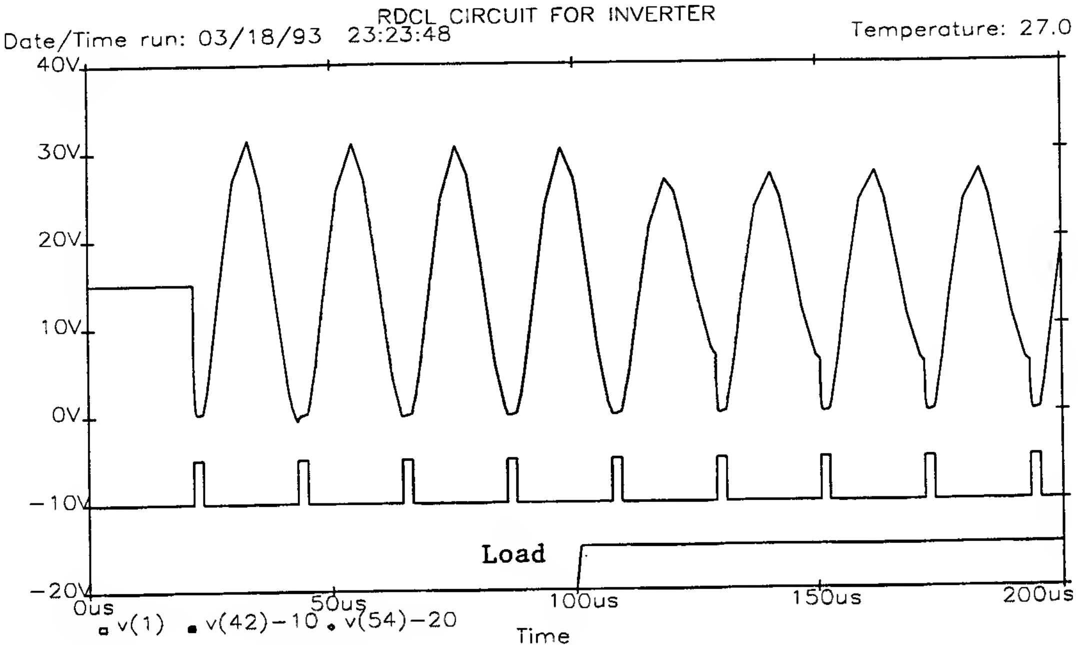


Fig. 18 simulated waveforms of link without current feedback

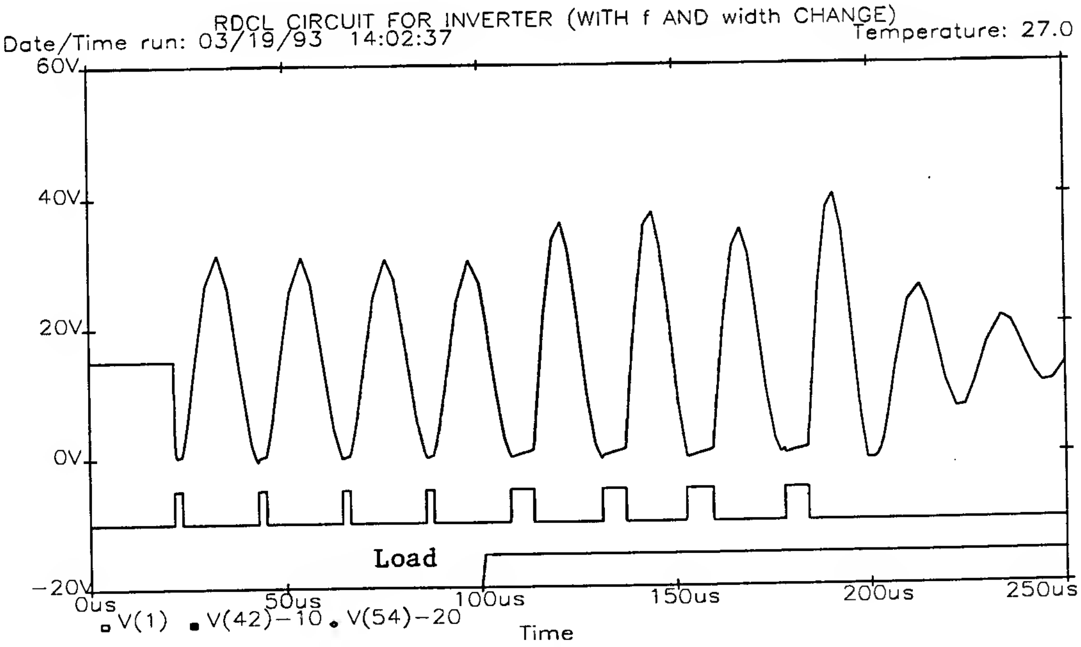


Fig. 19 Simulated waveforms of link with current feedback

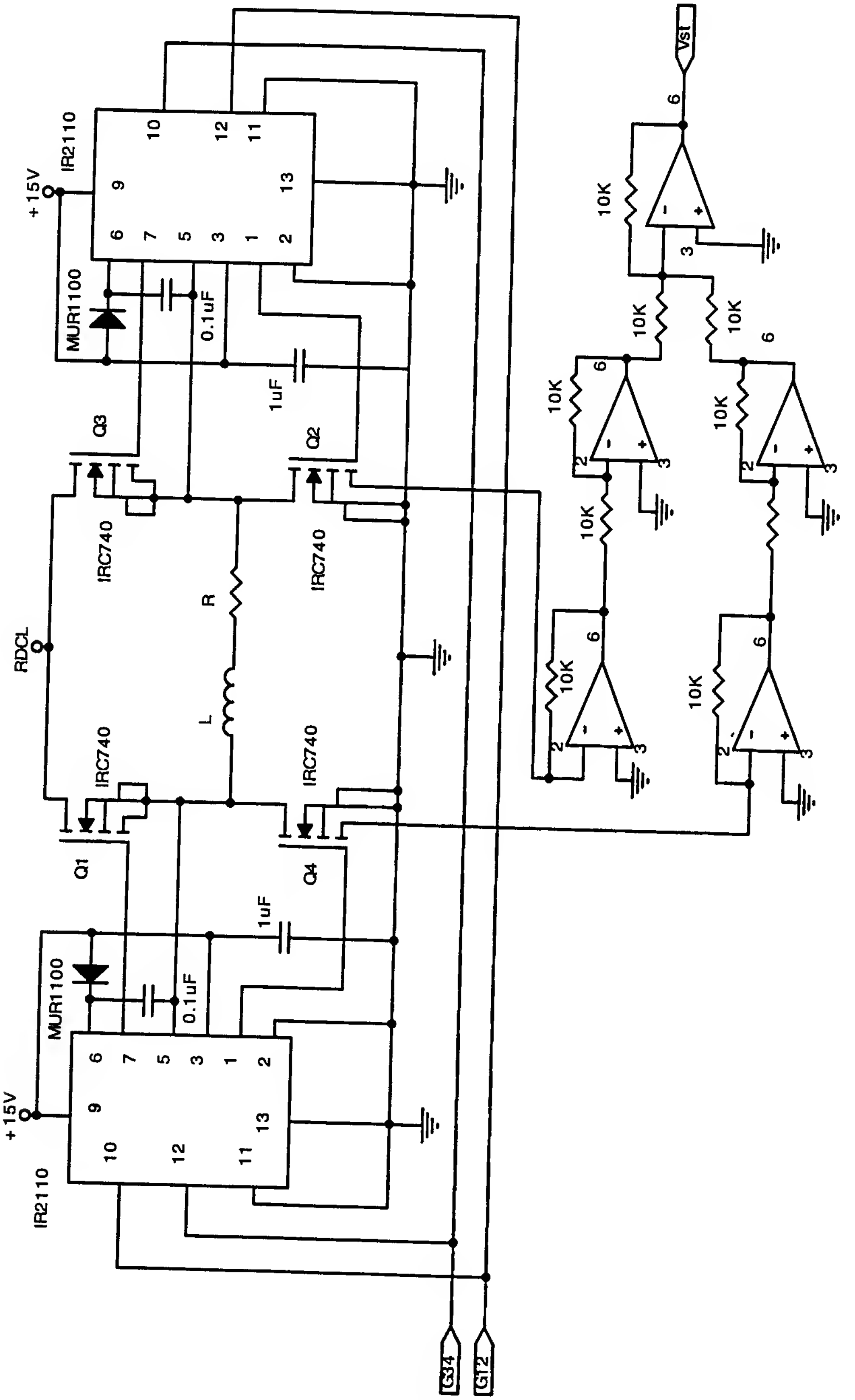


Fig. 20 Power circuit and current sensing circuit of RDCL inverter

To ensure zero-voltage switching, especially for resistive load, a new link-switch control circuit was developed. The value of $V_{min}$ at different loads was reduced to zero by feeding the current-sense output $V_{st}$ provided by the MOSFETs. The block diagram of the proposed control circuit is shown in Fig. 21. The strategy is to modify the frequency and duty cycle of the link-switch gate signal. The control circuit includes a 555 Timer used as a voltage-controlled oscillator, and a monostable providing a variable pulse-width output signal. Since the modifications in gate signal are to follow any change in load, the SPWM gate signal is also combined with $V_{st}$ in deciding the frequency and pulse width of the gate signal.

The reference signals that decide the frequency and pulse width of the gate signal under open loop are $V_{cf}$ and $V_{cp}$ respectively. The frequency reference signal $V_{cf}$, the total current sense signal $V_{st}$, and the SPWM signal are summed up to give the control signal to the 555 Timer. In the same way, the pulse width reference signal $V_{cp}$, the current sense signal $V_{st}$, and the SPWM signal are summed up to give the input to the monostable. The SPWM signal is used primarily to overcome the noise in the current sense output during the period when the inverter switches are off. The outputs from the 555 Timer and the monostable are combined using and 'OR' gate whose output signal has a variable frequency and variable duty cycle. The output of the OR gate is applied to the link-shorting MOSFET through a driver.

The modified RDCL inverter with current feedback was built and tested. The link-switch gate signal has a period of 20.5 μs and a width of 2 μs under no load. Fig. 22 shows the waveforms of link-voltage and link-switch gate signal with current feedback. Both the frequency and duty cycle change with load in order to ensure zero-crossing. The power loss in the inverter under different conditions were computed. For an output of 95 W, the power loss in a hard-switched inverter was measured as 9.1 W. The power losses in an ordinary RDCL inverter and the one with current feedback were measured as 5.4 W and 3.4 W respectively. There is an improvement in the efficiency with the addition of current feedback. It can be observed that the amplitude of the resulting link-voltage slightly increases with the addition of current feedback. If the increase in amplitude is objectionable, then the problem can be separately handled by adding a clamping circuit [2].

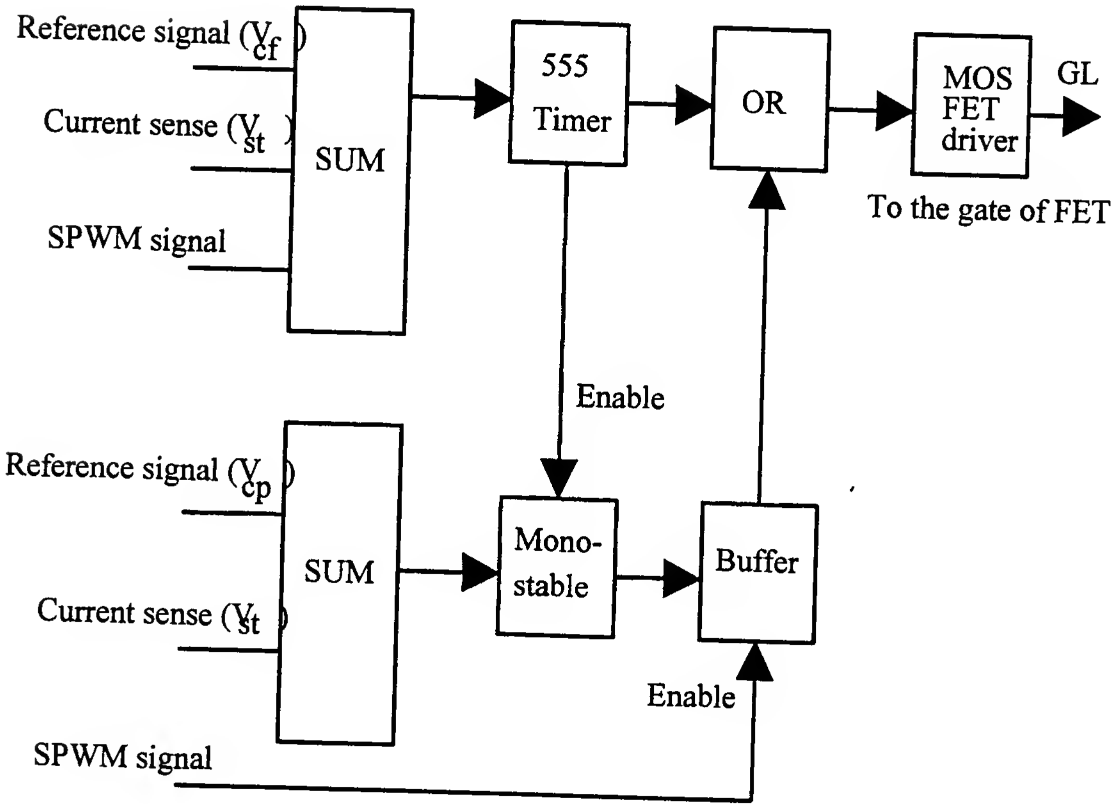
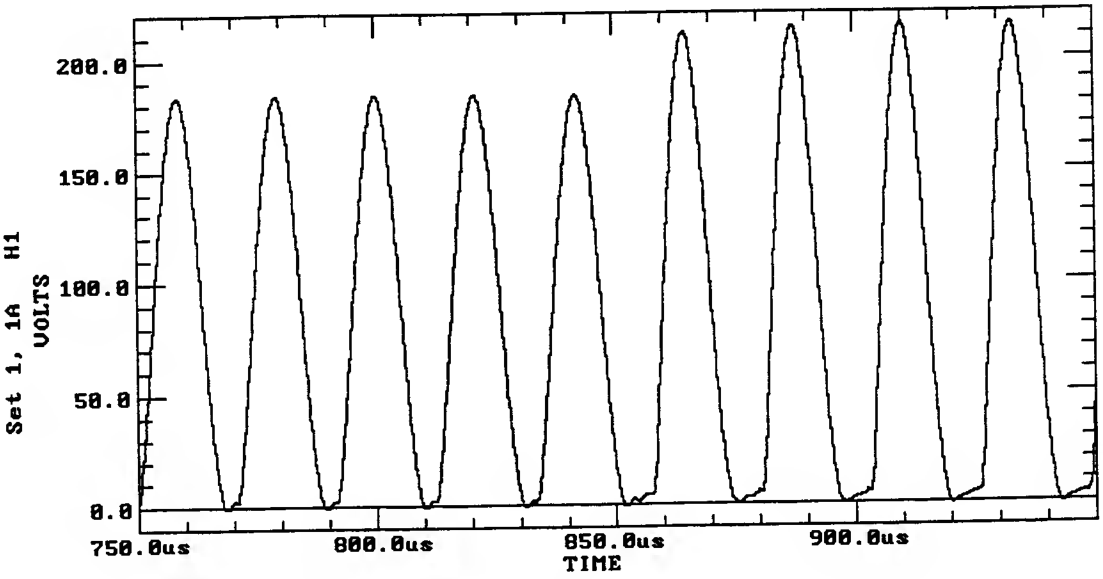
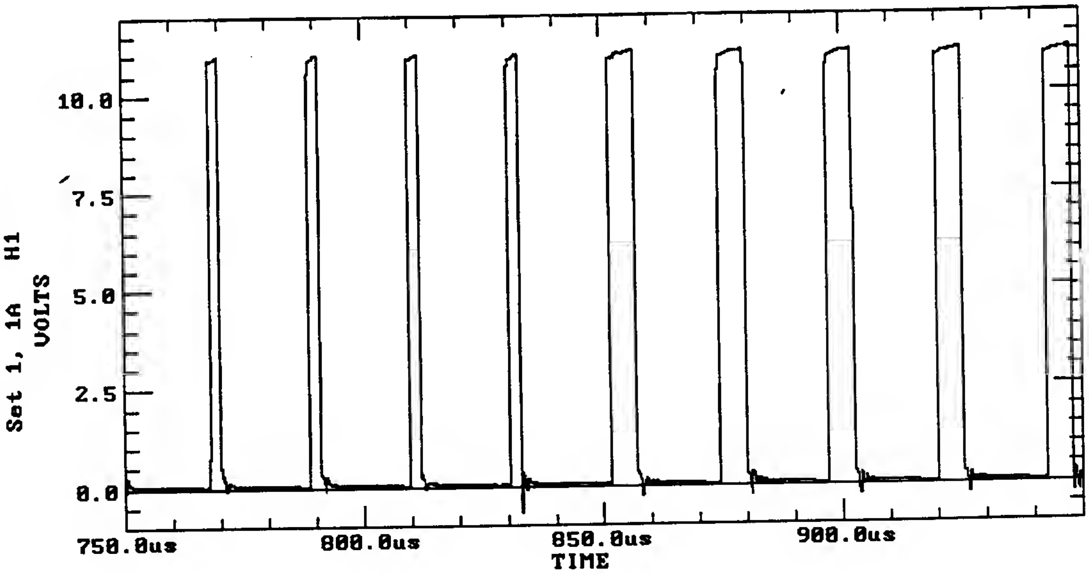


Fig. 21 Block diagram of signal prcessor with current feedback



(a) Link voltage



(b) Link-switch gate signal

Fig. 22 Experimental waveforms of link with current feedback

## 9. CONCLUSIONS

A single-phase and a three-phase resonant dc link inverters were developed and built primarily to study the performance and to investigate methods of improving it. The units developed consist of a resonant link section and an inverter section. The performance studies carried out include the measurement of power loss in the different sections. It was observed that a failure in zero-crossing results in additional power loss in the link-shorting Power MOSFET. A method of ensuring definite zero-crossings through the use of current feedback is described in the project. The possibility of providing load current feedback to ensure zero-voltage crossing was studied with the help of computer simulation. An RDCL inverter with current feedback was built using current sensing Power MOSFETs. The details of obtaining the current sense signal and generating the variable-frequency variable-duty-cycle gate signal for the link-shorting MOSFET are presented in the report. Several experimental waveforms obtained on the RDCL inverter are also presented.

## ACKNOWLEDGMENT

The author wishes to express his sincere thanks to the Air Force Office of Scientific Research, Bolling Air Force Base, Washington, DC for awarding the research initiation grant, and the authorities of North Dakota State University and the Electrical Engineering Department for providing the matching funds and the facilities. He is thankful to Mr. Joseph Weimer, Technical Area Manager, Electrical Technology Laboratory, Wright Patterson Air Force Base for providing the necessary facilities and encouragement during the initial investigations and the continued support thorough out the completion of the project. The author also wishes to thank Mr. Greg Fronista, Project Engineer, Wright Patterson Air Force, for providing the author with several useful informations during the execution of the project.

## REFERENCES

[1] D.M. Divan, "The resonant dc link converters - A new concept in static power conversion," IEEE Trans. Industry Applications, Vol. 25, No. 2, pp. 317-325, March/April 1989.

[2] D.M. Divan and G. Skibinski, "Zero-switching-loss inverters for high-power applications," IEEE Trans. Industry Applications, Vol. 25, No. 4, pp. 634-643, July/August 1989.

[3] K.S. Rajasekhara et al., "Resonant dc link inverter-fed ac machines control," Proc. of Power Electronics Specialists Conference, pp. 491-496, 1987.

[4] D.M. Divan et al., "Design methodologies for soft-switched inverters," Conf. Record of IEEE IAS Annual Meeting, pp. 758-766, 1987.

[5] PSPICE User's Manual, MicroSim Corporation, 1991

[6] J. Lai and B.K. Bose, "An induction motor drive using an improved high efficiency resonant dc link inverter," IEEE Trans. Power Electronics, Vol. 6, No. 3, pp. 504-513, July 1991.

[7] I. Smit et al., "Investigation of limitations in large converters with resonant input link above 30 kHz using GTOs," Proc. of Power Electronics Specialists Conference, pp. 1003-1009, 1989.

[8] M.H. Rashid, *Power Electronics,* 2nd Ed., Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] D.A. Bell, *Solid State Pulse Circuits,* 4th Ed., Prentice Hall , Englewood Cliffs, NJ, 1992.

[10] MBT (IGBT) Driver, Fuji Electric, 1990.

[11] VU-POINT: A digital data processing system, Maxwell Laboratories, 1991.

[12] Power Field Effect Transistor with current sensing capability, Technical Data, AD 11419, Motorola Semiconductors Inc., 1987.